ISMIR 2022 Tutorial • 4 December 2022

Deep learning for automatic mixing

Christian J. Steinmetz[1] Soumya Sai Vanka[1]

Gary Bromham[1] Marco A. Martínez Ramírez[2]

[1] Centre for Digital Music, Queen Mary University of London

[2] Sony Group Corporation, Tokyo, Japan

Presenters

Christian J. Steinmetz

Gary Bromham

Soumya Sai Vanka

Marco A. Martínez-Ramírez

Outline

Part 0	**Introduction**	Christian	*5 min*	1.5 hr
Part 1	**Audio Engineering**	Marco & Gary	*40 min*	
Part 2	**Automatic Mixing**	Christian	*40 min*	
	Break		*15 min*	
Part 3	**Implementation**	Soumya	*40 min*	1.5 hr
Part 4	**Evaluation**	Marco	*35 min*	
Part 5	**Conclusion**	Christian & Soumya	*15 min*	

More people are creating **audio** content

Podcasts

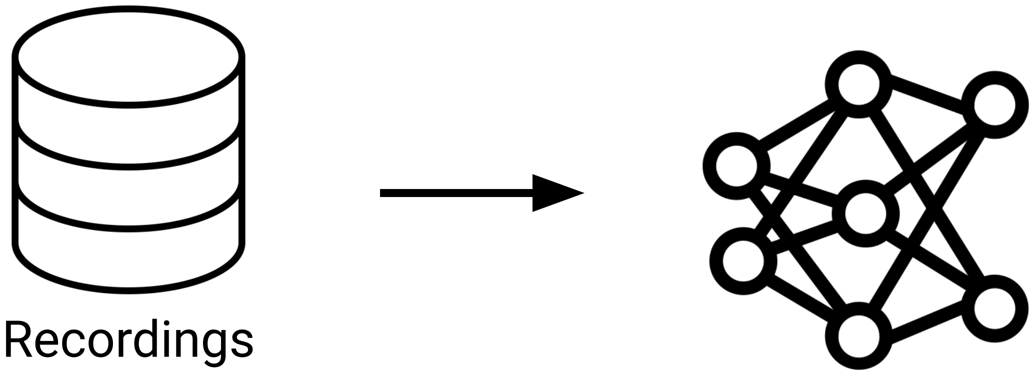
Sound for Video

Music

Short-form content

Demand for **high quality audio**

Producing **high quality audio** requires expertise



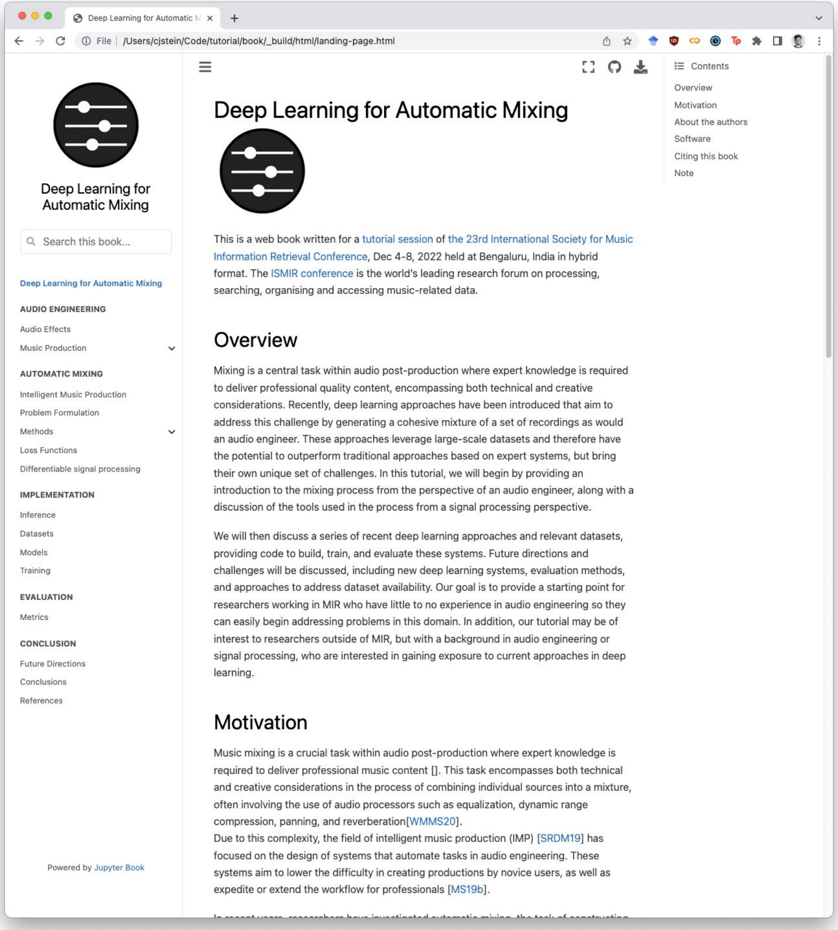
*Can we **learn** to produce recordings directly data?*

Goals

1. What is mixing and what should we consider for automix systems?
2. Framework for understanding and designing automix systems
3. Technical understanding of **two deep learning automix** models
4. How to **implement**, **train**, and **evaluate** these models
5. Ideas for future research directions

Book

https://dl4am.github.io/tutorial



Part 1

Audio Engineering

Gary Bromham

Marco A. Martínez-Ramírez

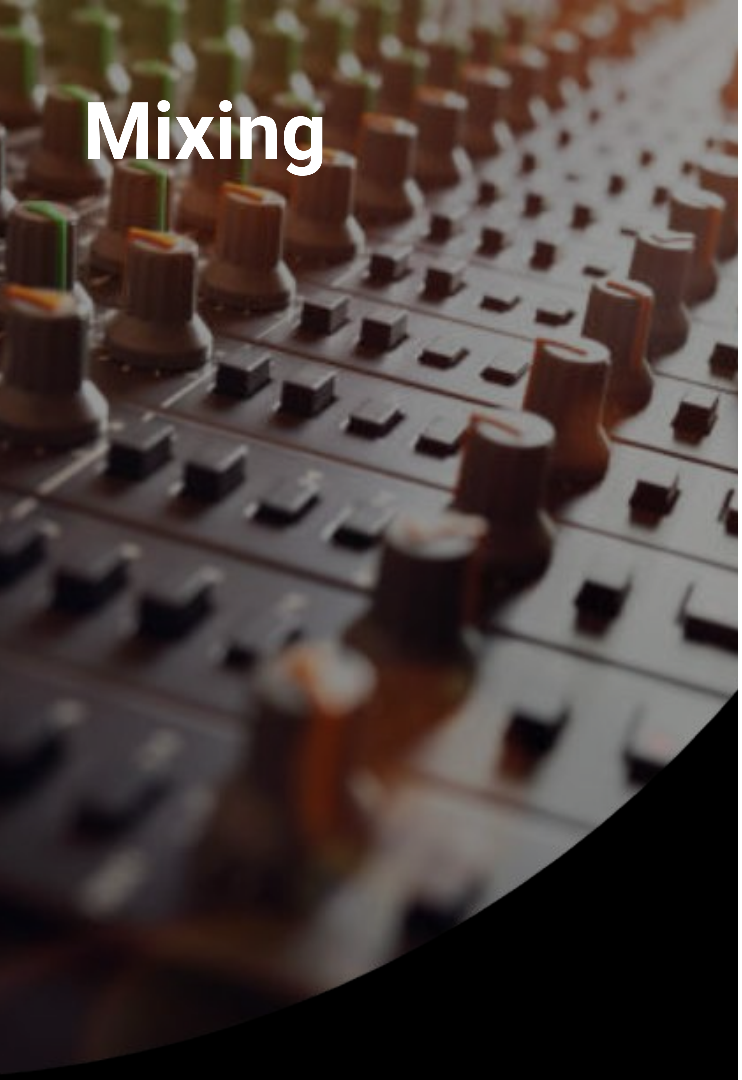
Levels

Music Production

Music production is a multi-dimensional creative process

It defines the life cycle of a piece of music

- Composition
- Recording
- Editing
- Mixing
- Mastering



Mixing

Audio mixing is the process of blending multitrack recordings

- Technical considerations together with creative, artistic or aesthetic decisions

Achieved with audio effects

- Gain
- Panning
- Equalization (EQ)
- Dynamic range compression (DRC)
- Artificial reverberation

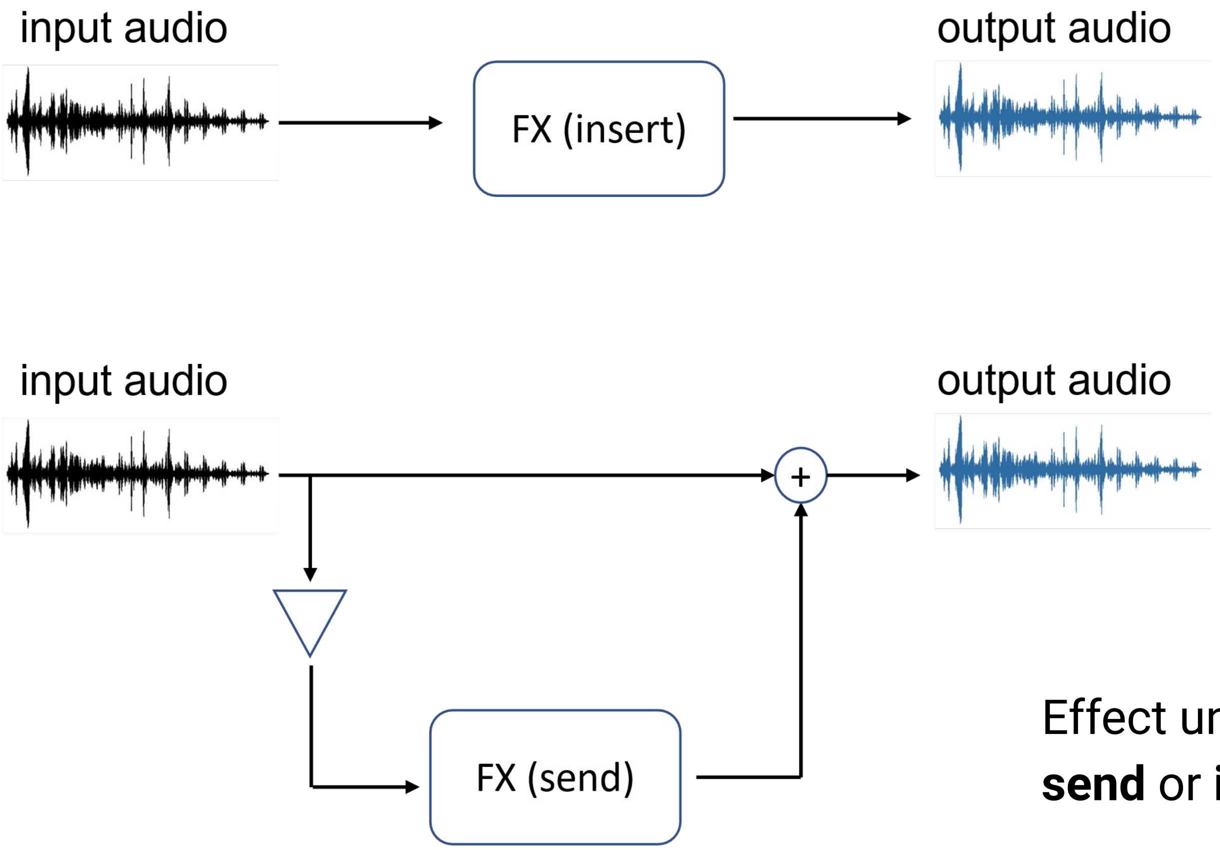
Audio effects are widely used

- Music
- Live performances
- Podcasts
- Films
- Games

To manipulate sounds

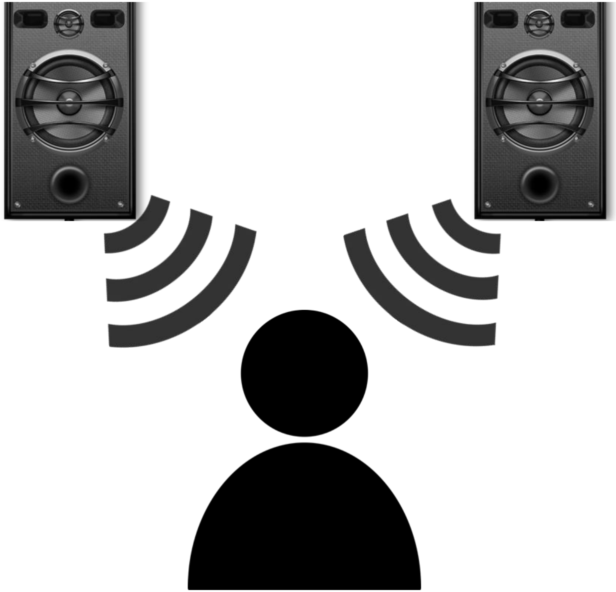
- Dynamics
- Frequency content
- Spatialisation
- Timbre


DELAY
DISTORTION
CHORUS
COMPRESSION
SATURATION
EQ
PHASER
LIMITER
REVERB
TREMOLO
VOCODER



Effect units can be applied as **send** or **insert** effects

Panning



Stereo panning is the positioning of sound sources using gain amplitude techniques that create azimuthal cues from mono sources

Panning

- Implemented **according to specific panning laws** which operate within a $\pi/2$ range
- Left and right speakers are at 0 and $\pi/2$, respectively
- The range of the **panning value θ is defined as $\theta \in [0, \pi/2]$**

Panning laws

- Linear panning
- Constant power panning
- -4.5 dB panning

Panning laws

Linear panning

- The gains of the left and right channels, $L(\theta)$ and $R(\theta)$, sum to 1

$$L(\theta) + R(\theta) = 1,$$
$$L(\theta) = \frac{2}{\pi}(\frac{\pi}{2} - \theta),$$
$$R(\theta) = \frac{2}{\pi}\theta.$$

Constant power panning

- The total power remains constant across all panning positions;

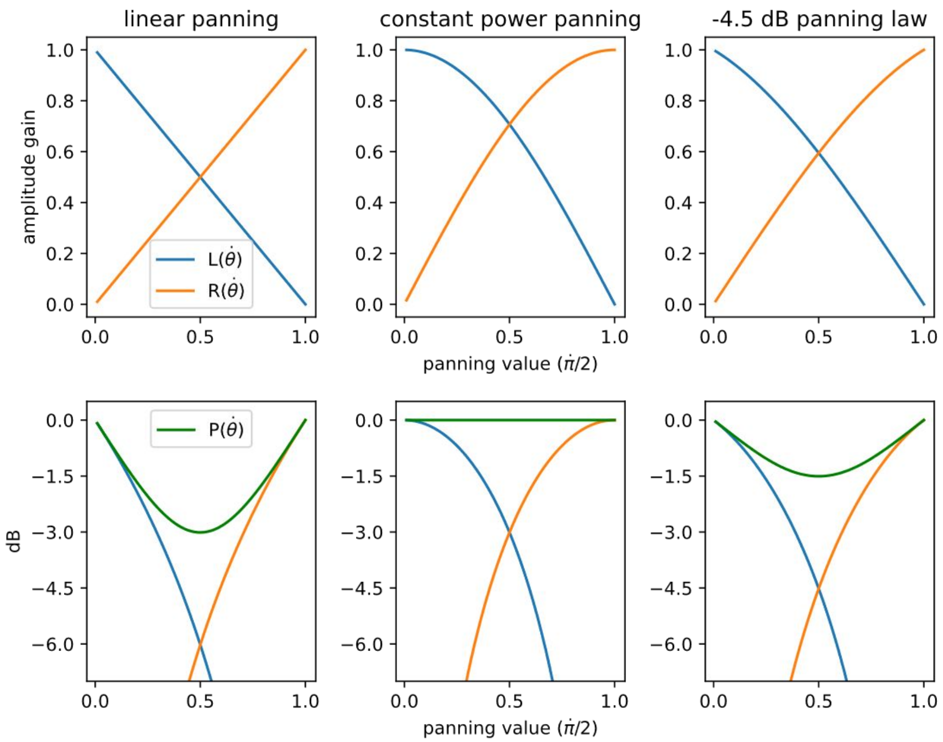
$$P(\theta) = L(\theta)^2 + R(\theta)^2$$
$$L(\theta) = \cos(\theta),$$
$$R(\theta) = \sin(\theta).$$

-4.5 dB panning

- Motivated for equal loudness panning, it is the square root of the product of the linear and constant power laws

$$L(\theta) = \sqrt{\frac{2}{\pi}(\frac{\pi}{2} - \theta) \cdot \cos(\theta)},$$
$$R(\theta) = \sqrt{\frac{2}{\pi}\theta \cdot \sin(\theta)}.$$

Panning laws



Equalization

EQ is the process of altering or adjusting the amplitude of various frequencies of a sound

It is used for many reasons, such as a

- Corrective filter to reduce masking
- Creative tool to shape harmonic and timbral characteristics

Equalization

- **Implemented via a filter bank** whose coefficients are obtained from the designed cut-off frequency *fc* and quality factor *Q*

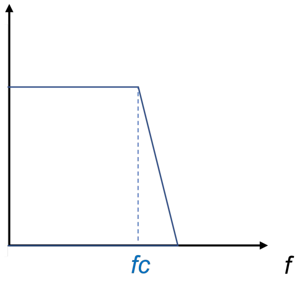
- The filter bank consists of Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filters, whose discrete difference equation is respectively:

$$y(n) = \sum_{k=0}^{M-1} a_k \cdot x(n-k), \qquad y(n) = \sum_{k=0}^{M-1} a_k \cdot x(n-k) - \sum_{k=1}^{N} b_k \cdot y(n-k).$$

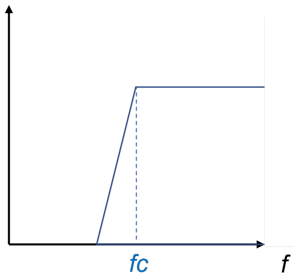
- Where *ak* and *bk* correspond to the *M* filter coefficients

Filter types

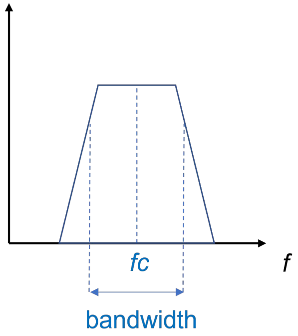
Lowpass



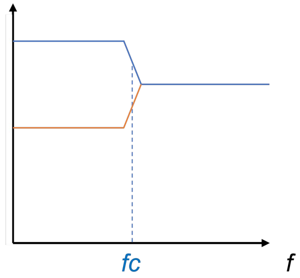
Highpass



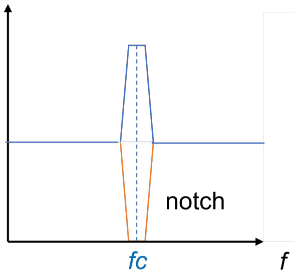
Bandpass



Shelving



Peaking



Compression

Compression is a nonlinear audio effect that is generally used to control the dynamic range of a sound

Extensively used by musicians and sound engineers

Nonlinear audio effects

- Nonlinear signal processing systems that **add harmonic or inharmonic frequency components** that are not present in the input signal
- This is known as a **harmonic and intermodulation distortion**
- Based on short term and long-term memory capabilities:
    - **Dynamic range processors** (DRC) such as compressors or limiters
    - **Distortion effects** such as tube amplifiers, fuzz distortion

Dynamic range processors

- The main purpose is to **change the variation in volume of the incoming audio**

- Apply a **time-varying** gain, which depends on an envelope follower along and waveshaping nonlinearity

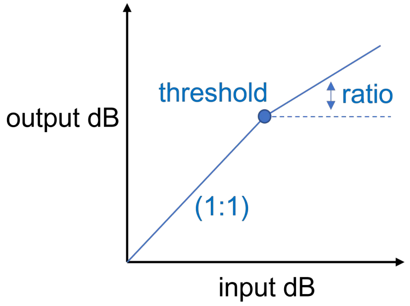
- This distorts the shape of the incoming waveform

- **Long-term memory**: the output depends on the current and previous samples

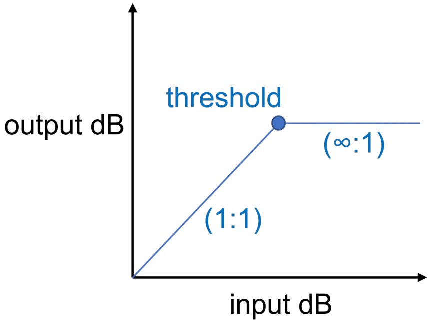
Parameters

- Threshold
- Ratio
- Attack and release times
- knee

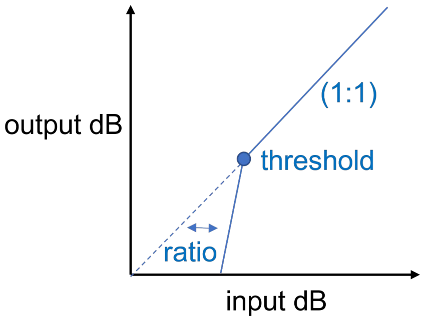
DRC types



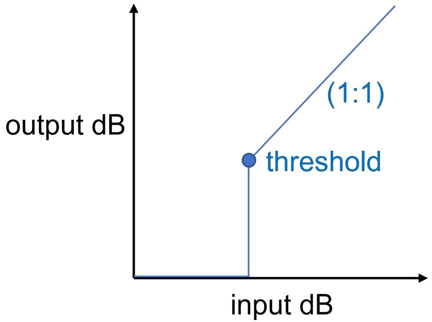
Compressor



Limiter



Expander



Noise Gate

DRC types

Multiband Compression

- Applies compression to selected frequency bands via a filter bank
- Each band is individually compressed

Sidechaining Compression

- The compressor has an additional input ("side input")
- The compressor is activated when the level of the input or side input is above the threshold

Artificial Reverberation

In the music and film industry, artificial reverberation was initially developed as a way of approximating acoustics of indoor spaces

This led to techniques that simulate reverberation, such as chamber, plate, spring and digital reverberators

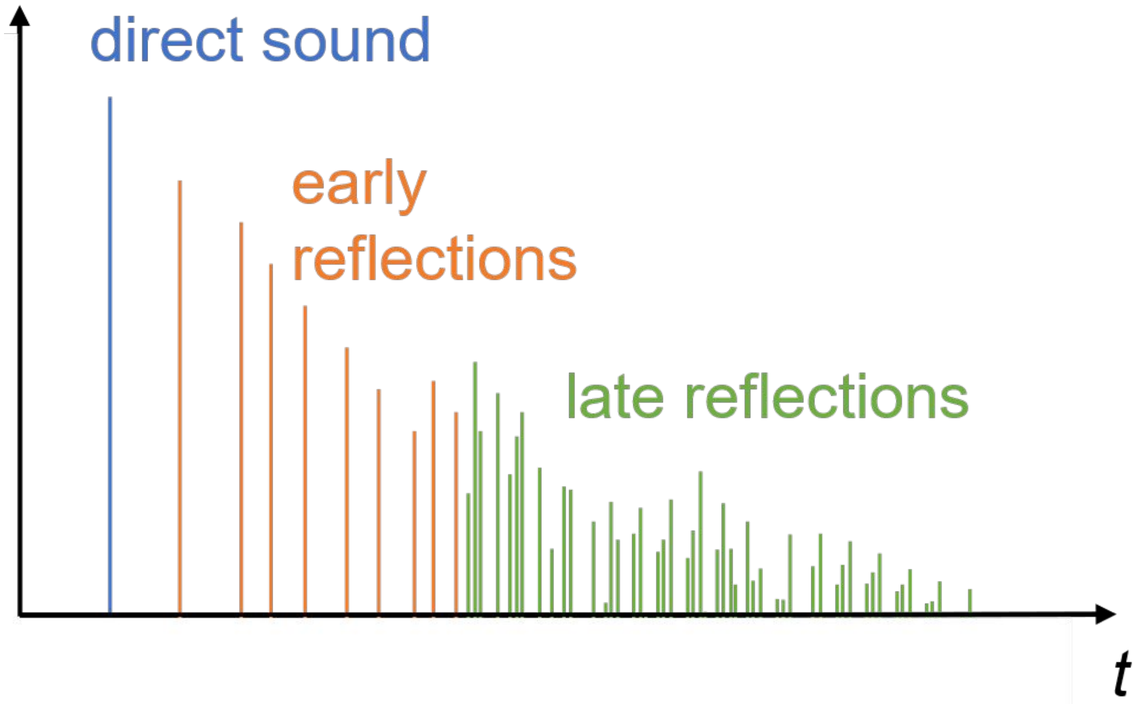
Artificial Reverberation

- It consists of **frequency-dependent reflections** of delayed and attenuated copies of the input or direct sound

- Each reflection is defined by the directivity of the sound source as well as the physical properties of the reflecting surfaces

- Reflections can be divided into: **direct sound, early reflections and late reflections**

Artificial Reverberation



Artificial Reverberation

- Most digital techniques **emulate the perceptual traits of impulse responses**

- Reverberation is approximately linear and time-invariant

- Methods rely on digital filters, delay networks and convolution-based algorithms

- **Types of artificial reverberation**

    - Comb and allpass filters
    - Feedback delay networks
    - Convolutional
    - Electromechanical

Artificial Reverberation

Comb and allpass filters

- Comb filters add a delay version of the input
    - Echoes that decay exponentially and are equally spaced in time: early reflections

$$y(n) = x(n - M) + gy(n - N),$$

- Allpass filters modify the phase relationships
    - Increases the overall echo density: late reflections

$$y(n) = x(n - M) - gx(n) + gy(n - M).$$

Artificial Reverberation

Convolutional

- Convolves the input signal with a recorded or estimated impulse response

Electromechanical

- ***Plate reverb*** is based on a large metal plate which vibrates due to a moving-coil transducer
  - Sound travels faster in metal than in air–this increases the echo density

- ***Spring reverb*** is based on helical springs suspended under low tension.
  - Spring vibrations results in an unusual combination of wave and dispersive propagation

Questions

Part 2

Automatic Mixing

Christian J. Steinmetz

Automatic Microphone Mixing*

DAN DUGAN

San Francisco, Calif. 94108

A method of analysis of sound reinforcement problems by means of active and passive speech zones is outlined. The need for automatic control of multimicrophone systems is defined, along with the problems associated with the use of voice-operated switches (VOX). Adaptive threshold gating is proposed as the best solution to the problem of active microphone detection. The development and performance of two effective automatic control systems is described.

A ZONAL THEORY OF SOUND REINFORCEMENT

A designer, engineer or contractor who works with sound equipment every day naturally tends to think only about the technical details when approaching a new problem. It is usual to start with deciding where to put the speakers and microphones, and what models will be best for the job. In most cases, this approach is completely valid. There is always a danger that our preoccupation with equipment and specifications will make us miss the real purpose of our efforts. A reinforcement system may have −1 dB frequency response and still not fill the needs of its users.

This paper describes some new inventions which promise to make the craft of sound reinforcement easier and more satisfying. Before getting into the details, I would like to make a short philosophical excursion into a sketch for a general theory of sound reinforcement. This theory is subject to much clarification and improvement.

Each person is the center of a zone in which he can communicate verbally. The size of this zone depends on the acoustical properties of the environment and on the person's ability as a speaker. The variables affecting the size of a person's speech zone may be tabulated:

1) effort
2) vocal ability
3) hearing acuity
4) ambient noise
5) reverberation.

Items 1) – 3) are human variables, 5) and 6) are environmental variables.

The border of this zone is not clearly defined, as all the variables change constantly, and the human ones are difficult to measure. If typical ranges of values are assigned to the variables, however, the design of environments will become possible in which speech will be relatively easy for almost all people, just as a door is designed to be high enough for people to pass without bumping their heads.

A frustrating thing about working in sound reinforcement is the lack of a direct and positive measurement of the effectiveness of communication transmitted through a system. The best available measurement is the articulation loss for consonants, AL_{cons} [2]. Measurement of AL_{cons} requires a group of observers whose responses can be treated statistically; this is too complex a procedure for daily use. AL_{cons} can be predicted from room data, but verification of these predictions is rare. Nevertheless, AL_{cons} is the best measurement available for speech transmission, and we will use the proposed 15% criterion.

* Presented May 14, 1975, at the Convention of the Audio Engineering Society, Los Angeles.

Dugan, 1975

Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, 15 September 2017

TEN YEARS OF AUTOMATIC MIXING

Brecht De Man and Joshua D. Reiss
Centre for Digital Music
Queen Mary University of London
{b.deman, joshua.reiss}@qmul.ac.uk

Ryan Stables
Digital Media Technology Lab
Birmingham City University
ryan.stables@bcu.ac.uk

ABSTRACT

Reflecting on a decade of Automatic Mixing systems for multitrack music processing, this paper positions the topic in the wider field of Intelligent Music Production, and seeks to motivate the existing and continued work in this area. Tendencies such as the introduction of machine learning and the increasing complexity of automated systems become apparent from examining a short history of relevant work, and several categories of applications are identified. Based on this systematic review, we highlight some promising directions for future research for the next ten years of Automatic Mixing.

1. MOTIVATION

The democratisation of audio technology has enabled music production on limited budgets, putting high-quality results within reach of anyone who has access to a laptop, a microphone and the abundance of free software on the web. Similarly, musicians are able to share their own content at very little cost and effort, again due to high availability of cheap technology. Despite this, a skilled mix engineer is often still needed in order to deliver professional-standard material. Raw, recorded tracks almost always require a considerable amount of processing before being ready for distribution, such as balancing, panning, equalisation (EQ), dynamic range compression and artificial reverberation, to name a few. Furthermore, an amateur music producer will

Meanwhile, professional audio engineers are often under pressure to produce high-quality content quickly and at low cost [3]. While they may be unlikely to relinquish control entirely to autonomous mix software, assistance with tedious, time-consuming tasks would be highly beneficial. This can be implemented via more powerful, intelligent, responsive, intuitive algorithms and interfaces [4].

Throughout the history of technology, innovation has traditionally been met with resistance and scepticism, in particular from professional users who fear seeing their roles disrupted or made obsolete. Music production technology may be especially susceptible to this kind of opposition, as it is characterised by a tendency towards nostalgia, skeuomorphisms and analogue workflows [1], and it is concerned with aesthetic value in addition to technical excellence and efficiency. However, the evolution of music is intrinsically linked to the development of new instruments and tools, and essentially utilitarian inventions such as automatic vocal riding, drum machines, electromechanical keyboards and digital pitch correction have been famously used and abused for creative effect. These advancements have changed the nature of the sound engineering profession from primarily technical to increasingly expressive. Generally, there is economic, technological and artistic merit in exploiting the immense computing power and flexibility that today's digital technology affords, to venture away from the rigid structure of the traditional music production toolset.

De Man et al., 2017

1. **Knowledge-based Systems**
   Gonzalez et al. 2007, De Man et al. 2013,

2. **Classical ML-based Systems**
   Scott and Kim, 2011

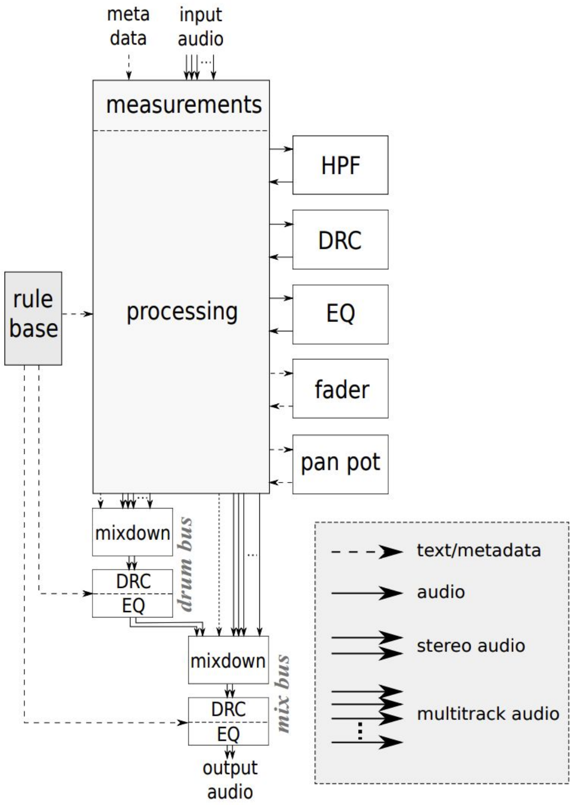
3. **Deep Learning-based Systems**
   Martinez Ramirez et al., 2021 and Steinmetz et al. 2020

Knowledge-based or **Expert systems**

Design a set of rules based to create a mix based on analysis of the inputs.

Pro: Explainable decisions

Con: Often lacks sufficient complexity



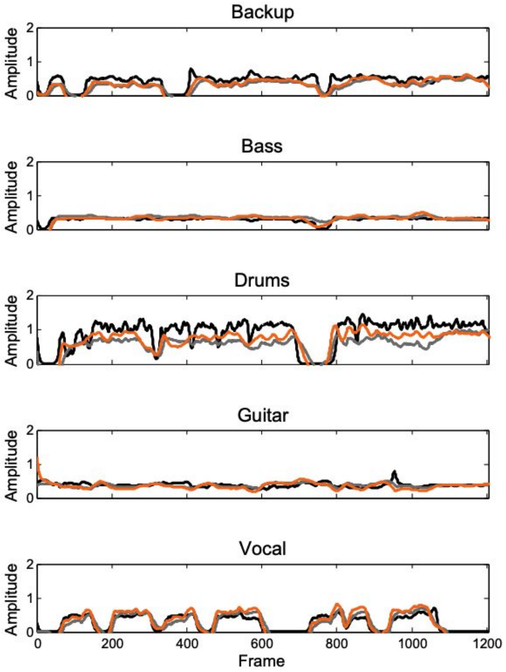
A knowledge-engineered autonomous mixing system
Brecht De Man, Joshua D. Reiss AES 2013

Machine Learning*

Learn to create a mix by leveraging parametric data collected from pros.

Pro: Greater model flexibility

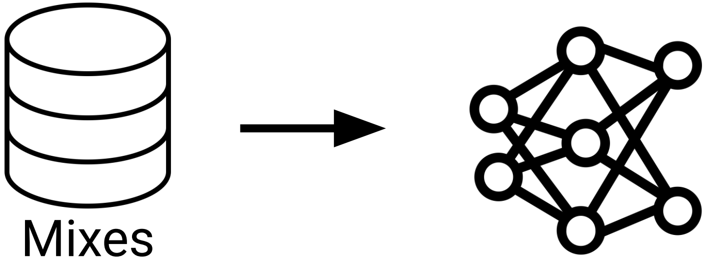
Con: Requires data (parametric)



*Approaches that use classical machine learning techniques

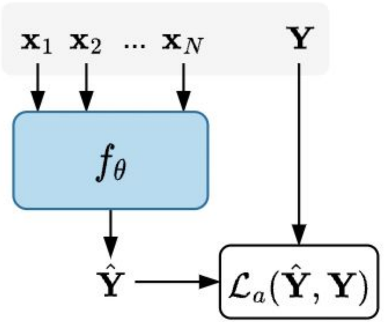
Analysis of acoustic features for automated multitrack mixing
Jeffrey J. Scott. Youngmoo E. Kim ISMIR 2011

Deep Learning

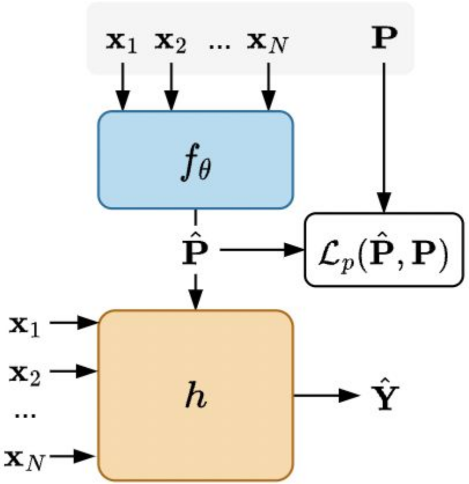


*Can we **learn** to produce mixes directly from data?*

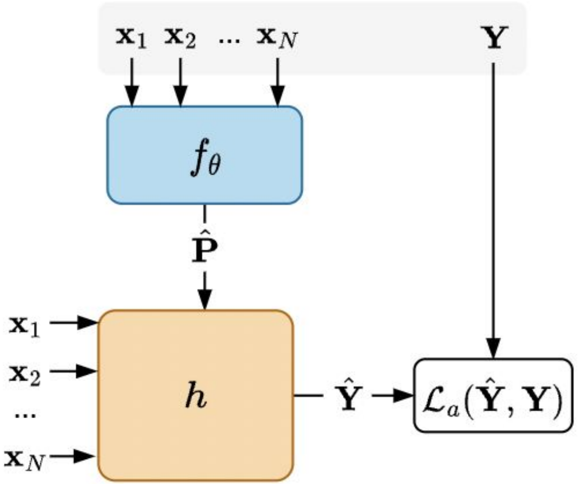
Problem Formulation



Direct Transformation

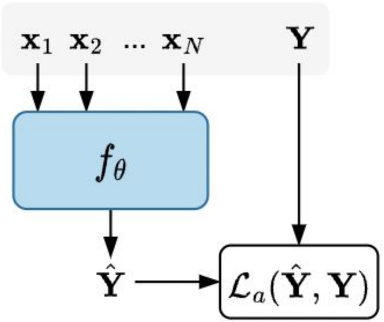


Parameter Estimation
(Parameter Loss)



Parameter Estimation
(Audio Loss)

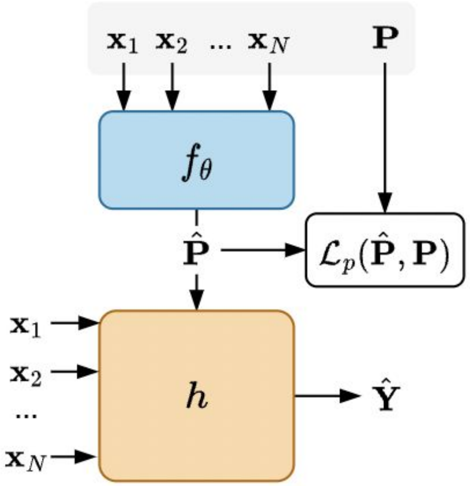
Direct Transformation



Direct Transformation

Parameter Estimation

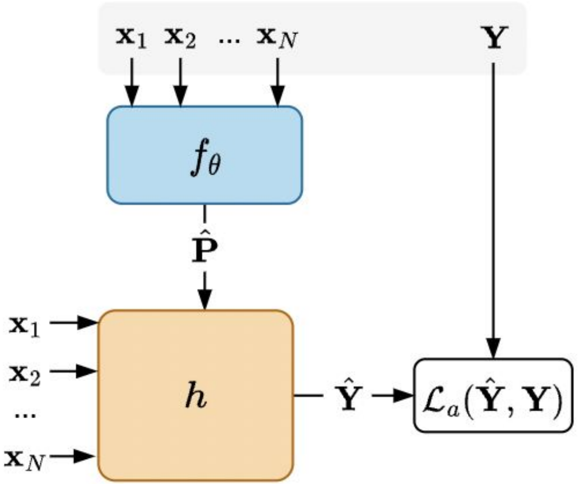
Parameter space loss



Parameter Estimation
(Parameter Loss)

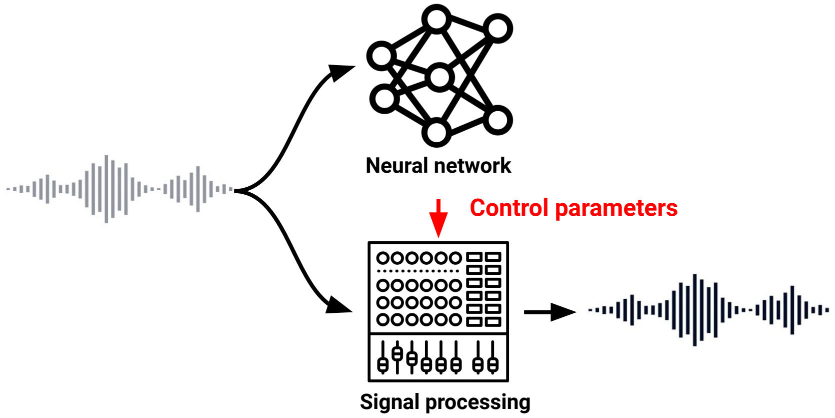
Parameter Estimation

Audio domain loss



Parameter Estimation
(Audio Loss)

Differentiable signal processing



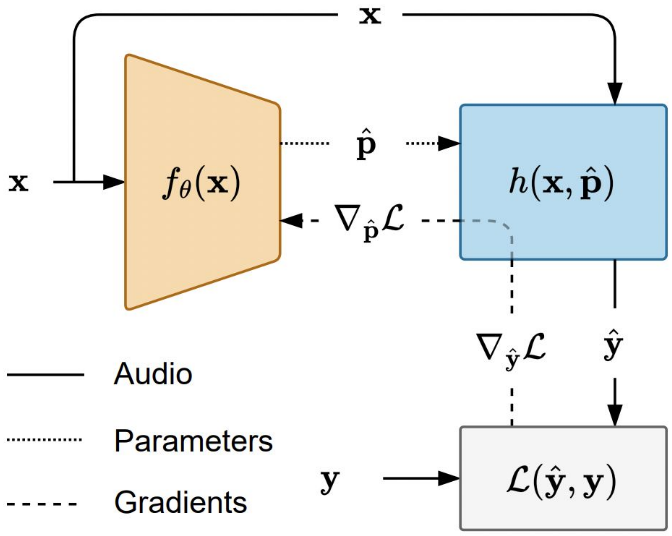
- Leveraging existing DSP tools and knowledge
- High quality audio processing with few artifacts
- Human understandable outputs that can be adjusted
- Efficient and can easily run in real-time on CPU

Differentiable signal processing

Non-differentiable

Discontinuous
(Discrete options)

Recursive operations



Backpropgation through the DSP is non-trivial

Techniques

1. **Automatic differentiation (AD)**
   Engel et al. 2020

2. **Neural proxies and hybrids (NP)**
   Steinmetz et al. 2020, Steinmetz et al. 2022

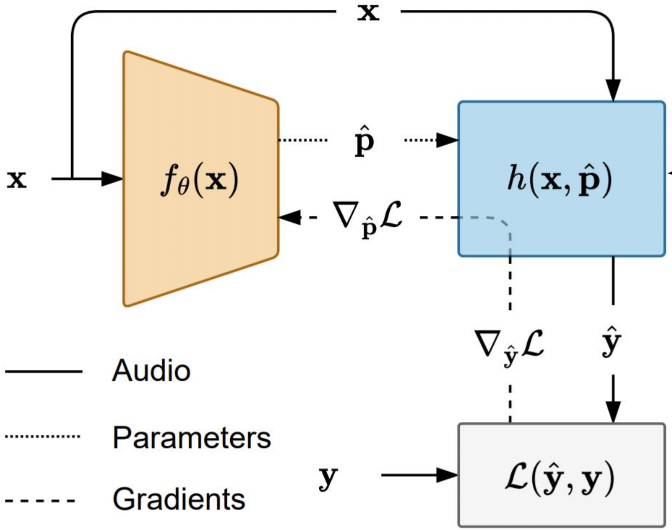
3. **Numerical gradient approximation (NGA)**
   Martínez Ramírez et al. 2021

Automatic differentiation

White-box

Requires hacks or tricks for each DSP

Doesn't work for all kinds of DSP




```
A = 10 ** (gain_dB / 40.0)
w0 = 2 * math.pi * (cutoff_freq / sample_rate)
alpha = torch.sin(w0) / (2 * q_factor)
cos_w0 = torch.cos(w0)
sqrt_A = torch.sqrt(A)

if filter_type == "high_shelf":
    b0 = A * ((A + 1) + (A - 1) * cos_w0 + 2 * sqrt_A * alpha)
    b1 = -2 * A * ((A - 1) + (A + 1) * cos_w0)
    b2 = A * ((A + 1) + (A - 1) * cos_w0 - 2 * sqrt_A * alpha)
    a0 = (A + 1) - (A - 1) * cos_w0 + 2 * sqrt_A * alpha
    a1 = 2 * ((A - 1) - (A + 1) * cos_w0)
    a2 = (A + 1) - (A - 1) * cos_w0 - 2 * sqrt_A * alpha
```

Explicitly define signal processing operations in autodiff framework

PyTorch TensorFlow

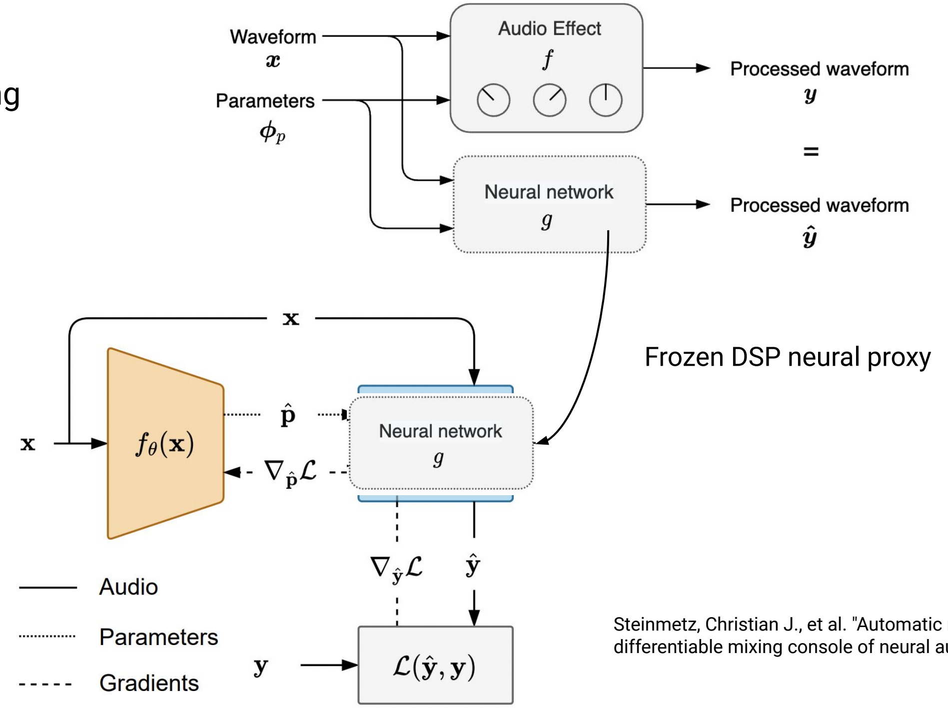
Engel, Jesse, et al. "DDSP: Differentiable digital signal processing." *ICLR* (2021).

Neural proxy

(1) Pretraining

(2) Training

(3) Inference

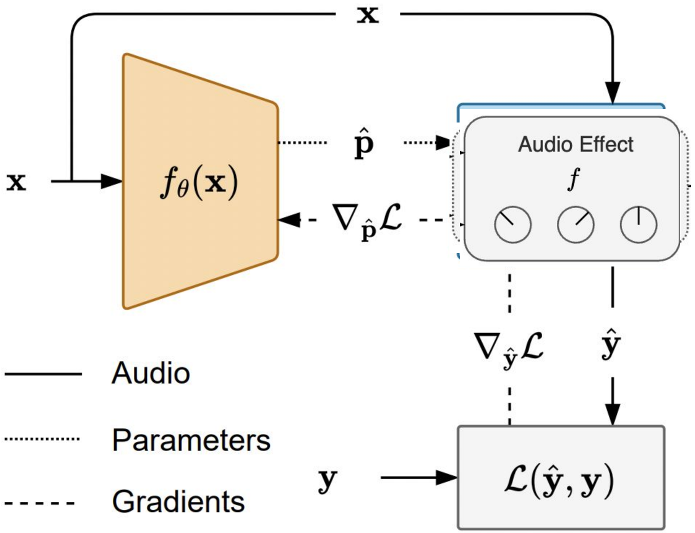


Steinmetz, Christian J., et al. "Automatic multitrack mixing with a differentiable mixing console of neural audio effects." ICASSP, 2021.

Neural proxy hybrid

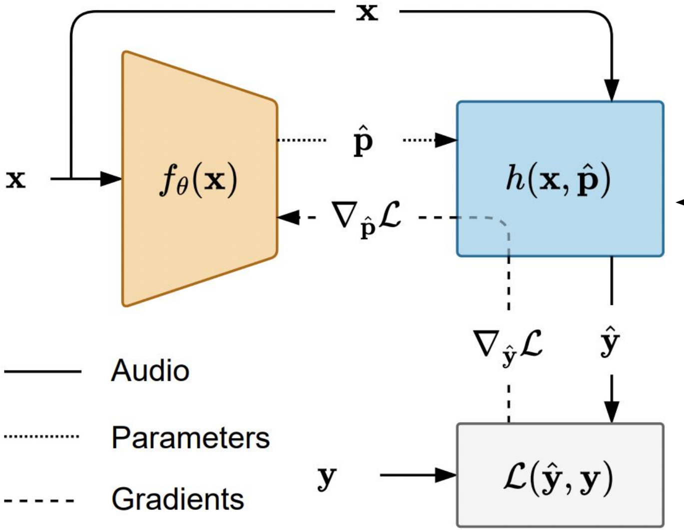
(2) Training

(3) Inference

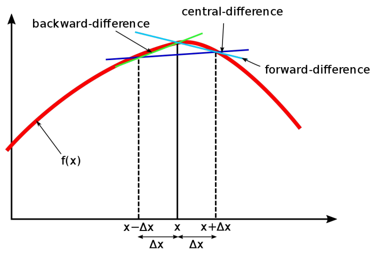


Use original DSP during inference

Gradient approximation



Finite differences (FD)



$$\frac{\hat{h}(x, p_i)}{p_i} = \frac{h(x, p + \varepsilon\Delta^P) - h(x, p - \varepsilon\Delta^P)}{2\varepsilon\Delta_i^P}, \qquad (2)$$

where ε is a small, non-zero value and $\Delta^P \in \mathbb{R}^P$ is a random vector sampled from a symmetric Bernoulli distribution ($\Delta_i^P = \pm 1$) [46].

Simultaneous perturbation stochastic approximation (SPSA)

Martínez Ramírez, Marco A., et al. "Differentiable signal processing with black-box audio effects." ICASSP, 2021.

Considerations

Interpretability

Input Taxonomy

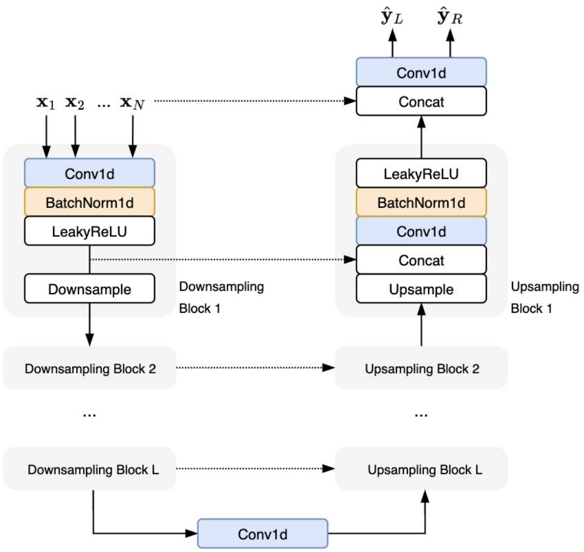
Controllability

Fidelity

Context

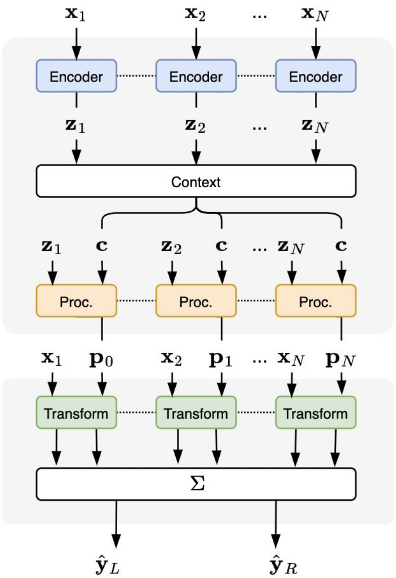
Expressivity

Deep Learning Models



Mix-Wave-U-Net

Direct Transformation



Differentiable Mixing Console

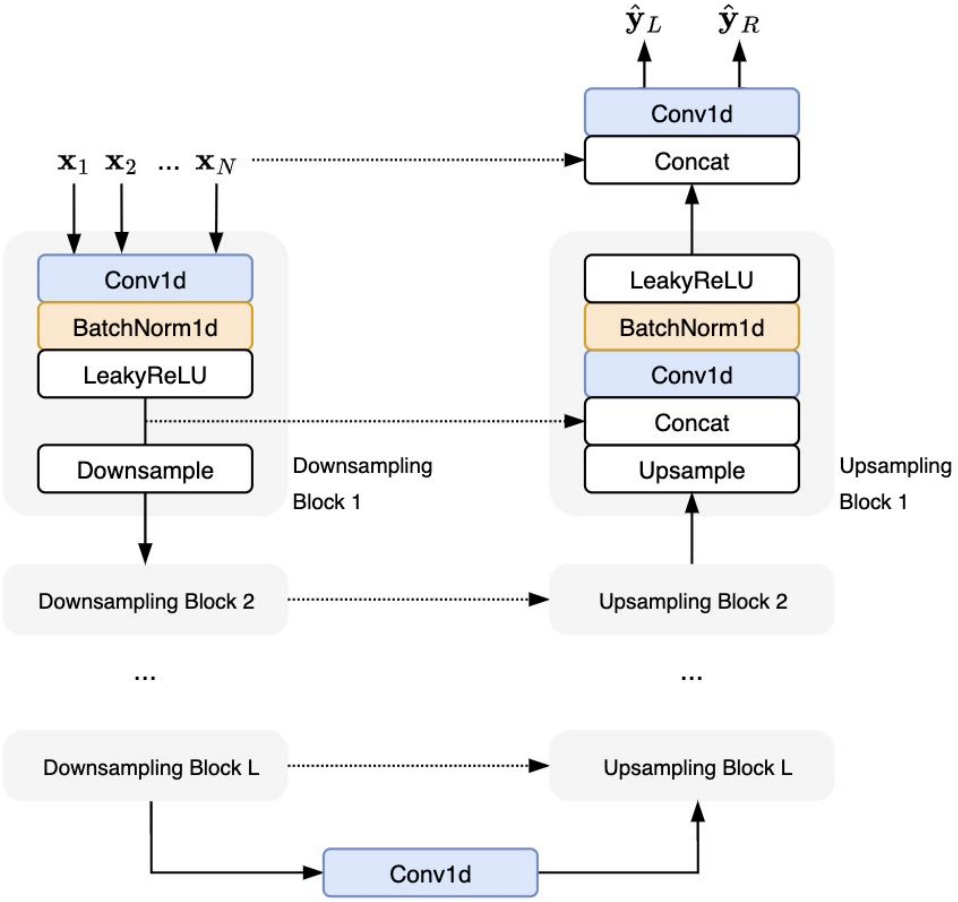
Parameter Estimation

“A Deep Learning Approach to Intelligent Drum Mixing with Wave-U-Net”, Martínez Ramírez et al. 2021

“Automatic Multitrack Mixing with a Differentiable Mixing Console of Neural Audio Effects”, Steinmetz et al. 2021

Mix-Wave-U-Net

Direct transformation



Mix-Wave-U-Net

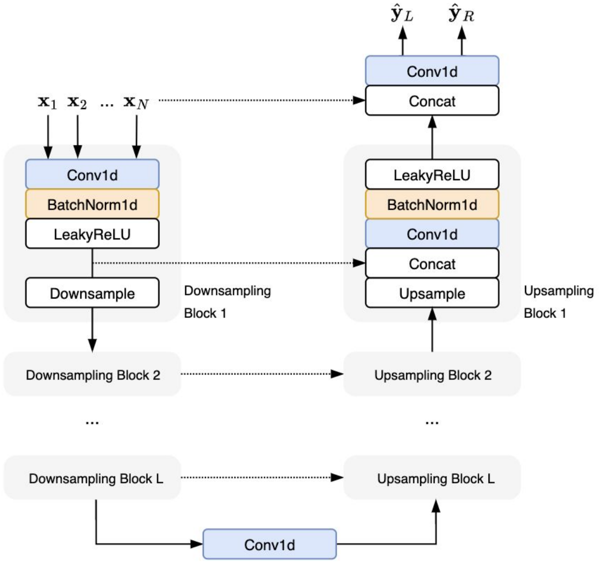
A Deep Learning Approach to Intelligent Drum Mixing with the Wave-U-Net

Marco A. Martínez Ramírez[1*], **Daniel Stoller**[1*], **AND David Moffat**[2], *AES Student Member*
(m.a.martinezramirez@qmul.ac.uk) (d.stoller@qmul.ac.uk) (david.moffat@plymouth.ac.uk)

[1]*Centre for Digital Music, Queen Mary University of London, London, United Kingdom*
[2]*University of Plymouth, Plymouth, United Kingdom*

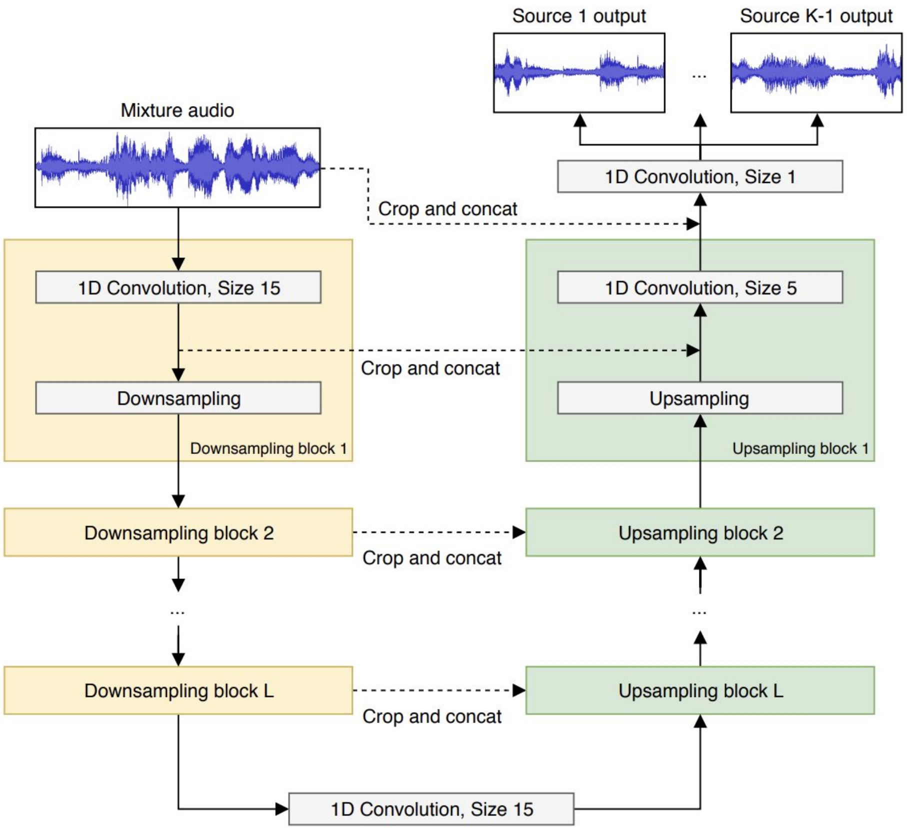
[*] *These authors contributed equally to this work.*

The development of intelligent music production tools has been of growing interest in recent years. Deep learning approaches have been shown as being a highly effective method for approximating individual audio effects. In this work, we propose an end-to-end deep neural network based on the Wave-U-Net to perform automatic mixing of drums. We follow an end-to-end approach, where raw audio from the individual drum recordings is the input of the system and the waveform of the stereo mix is the output. We compare the system to existing machine learning approaches to intelligent drum mixing. Through a subjective listening test, we explore the performance of these systems when processing various types of drum mixes. We report that the mixes generated by our model are virtually indistinguishable from professional human mixes, while also outperforming previous intelligent mixing approaches.



Wave-U-Net

Music source separation



Wave-U-Net

WAVE-U-NET: A MULTI-SCALE NEURAL NETWORK FOR END-TO-END AUDIO SOURCE SEPARATION

Daniel Stoller
Queen Mary University of London
d.stoller@qmul.ac.uk

Sebastian Ewert
Spotify
sewert@spotify.com

Simon Dixon
Queen Mary University of London
s.e.dixon@qmul.ac.uk

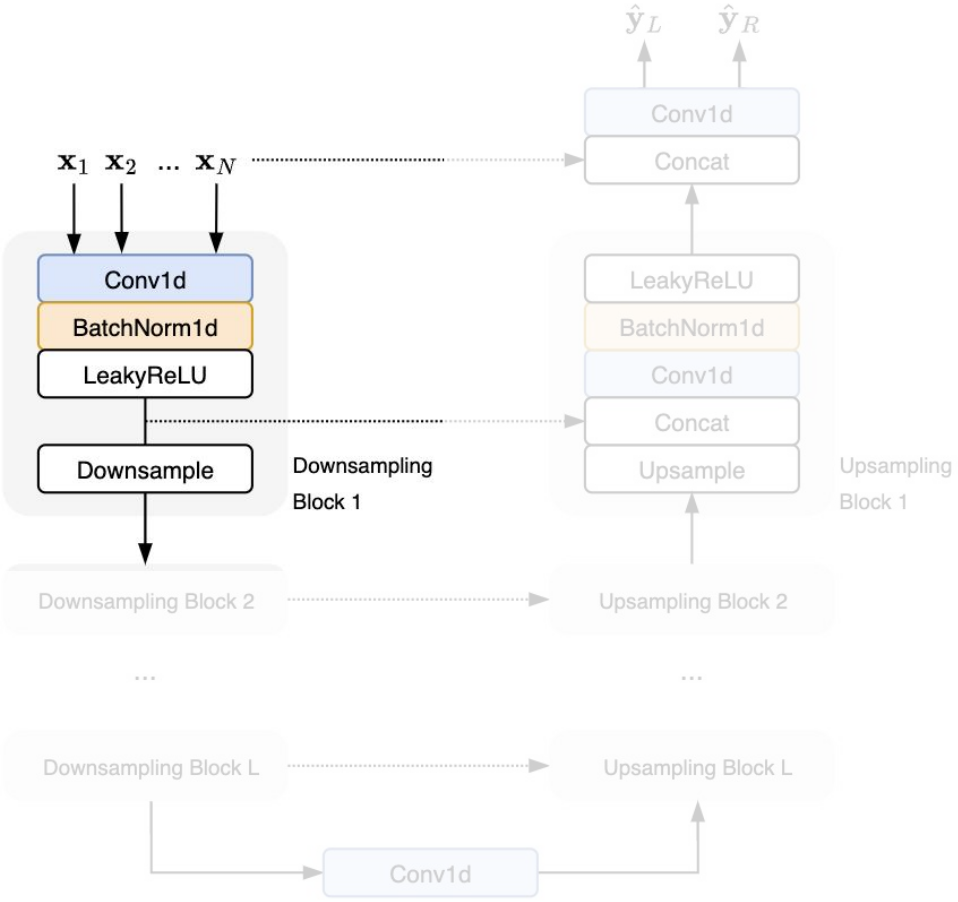
ABSTRACT

Models for audio source separation usually operate on the magnitude spectrum, which ignores phase information and makes separation performance dependant on hyper-parameters for the spectral front-end. Therefore, we investigate end-to-end source separation in the time-domain, which allows modelling phase information and avoids fixed spectral transformations. Due to high sampling rates for audio, employing a long temporal input context on the sample level is difficult, but required for high quality separation results because of long-range temporal correlations. In this context, we propose the Wave-U-Net, an adaptation of the U-Net to the one-dimensional time domain, which repeatedly resamples feature maps to compute and com-

This approach has several limitations. Firstly, the STFT output depends on many parameters, such as the size and overlap of audio frames, which can affect the time and frequency resolution. Ideally, these parameters should be optimised in conjunction with the parameters of the separation model to maximise performance for a particular separation task. In practice, however, the transform parameters are fixed to specific values. Secondly, since the separation model does not estimate the source phase, it is often assumed to be equal to the mixture phase, which is incorrect for overlapping partials. Alternatively, the Griffin-Lim algorithm can be applied to find an approximation to a signal whose magnitudes are equal to the estimated ones, but this is slow and often no such signal exists [8]. Lastly, the mixture phase is ignored in the estimation of sources.

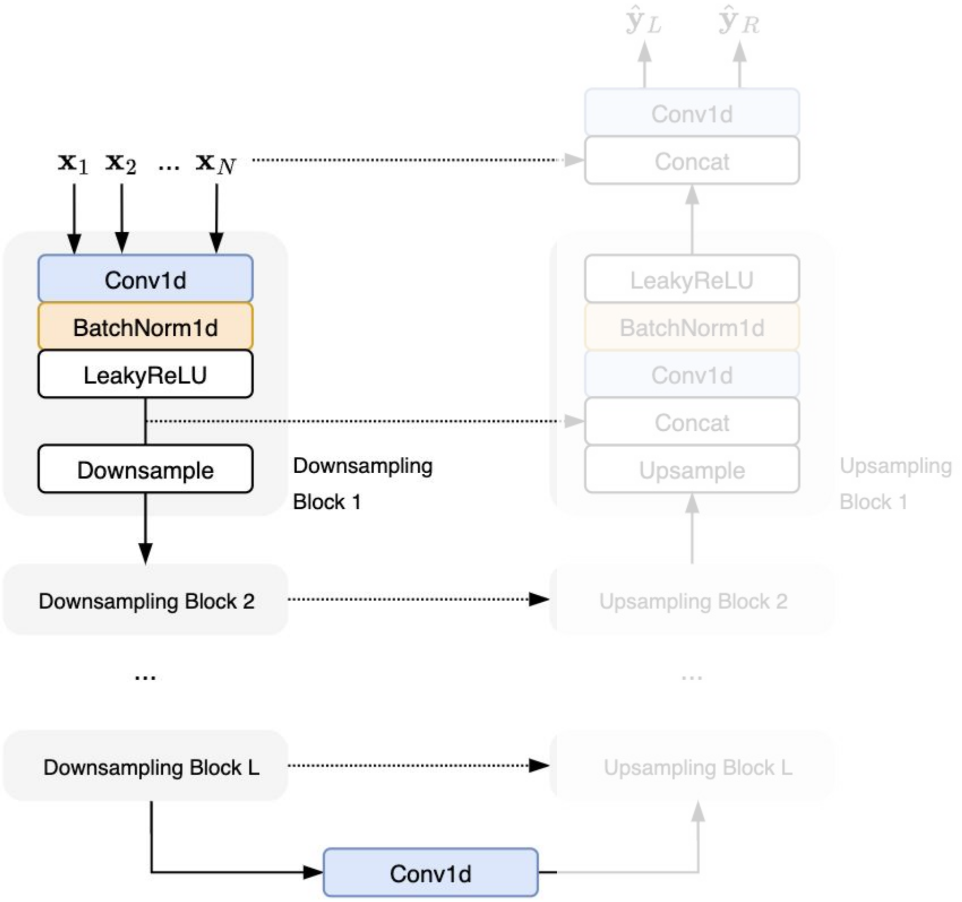
Mix-Wave-U-Net

Downsampling block



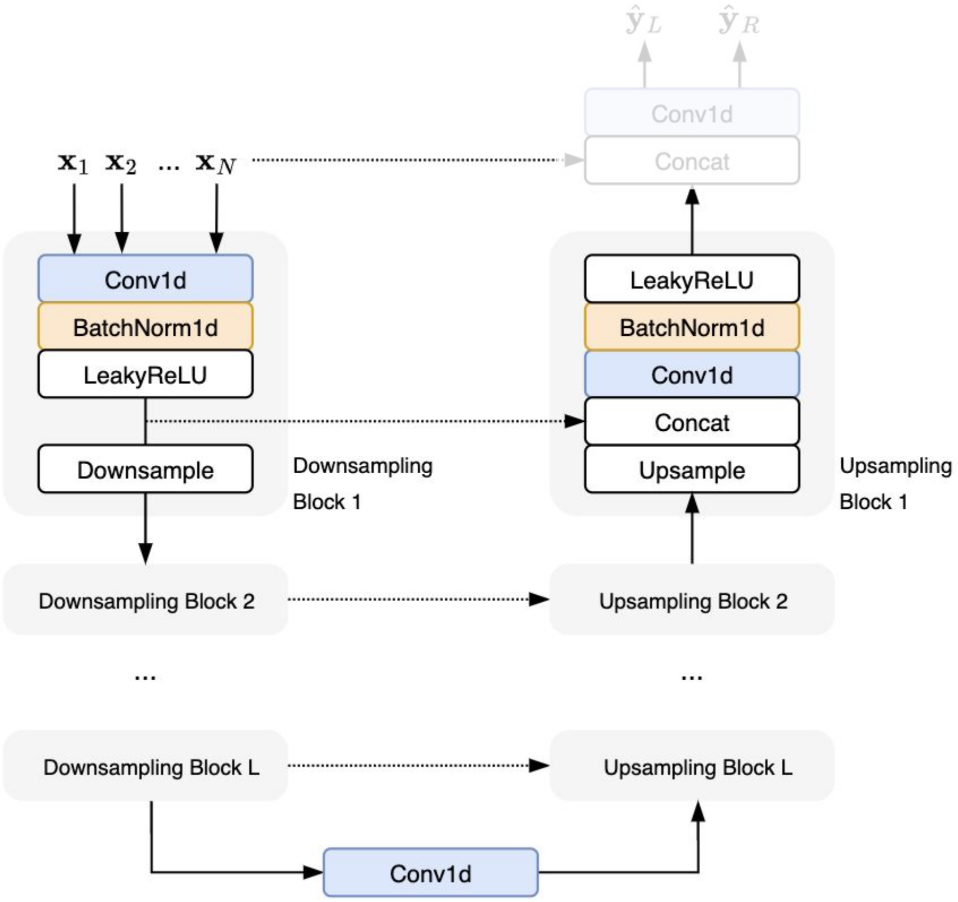
Mix-Wave-U-Net

Downsampling block

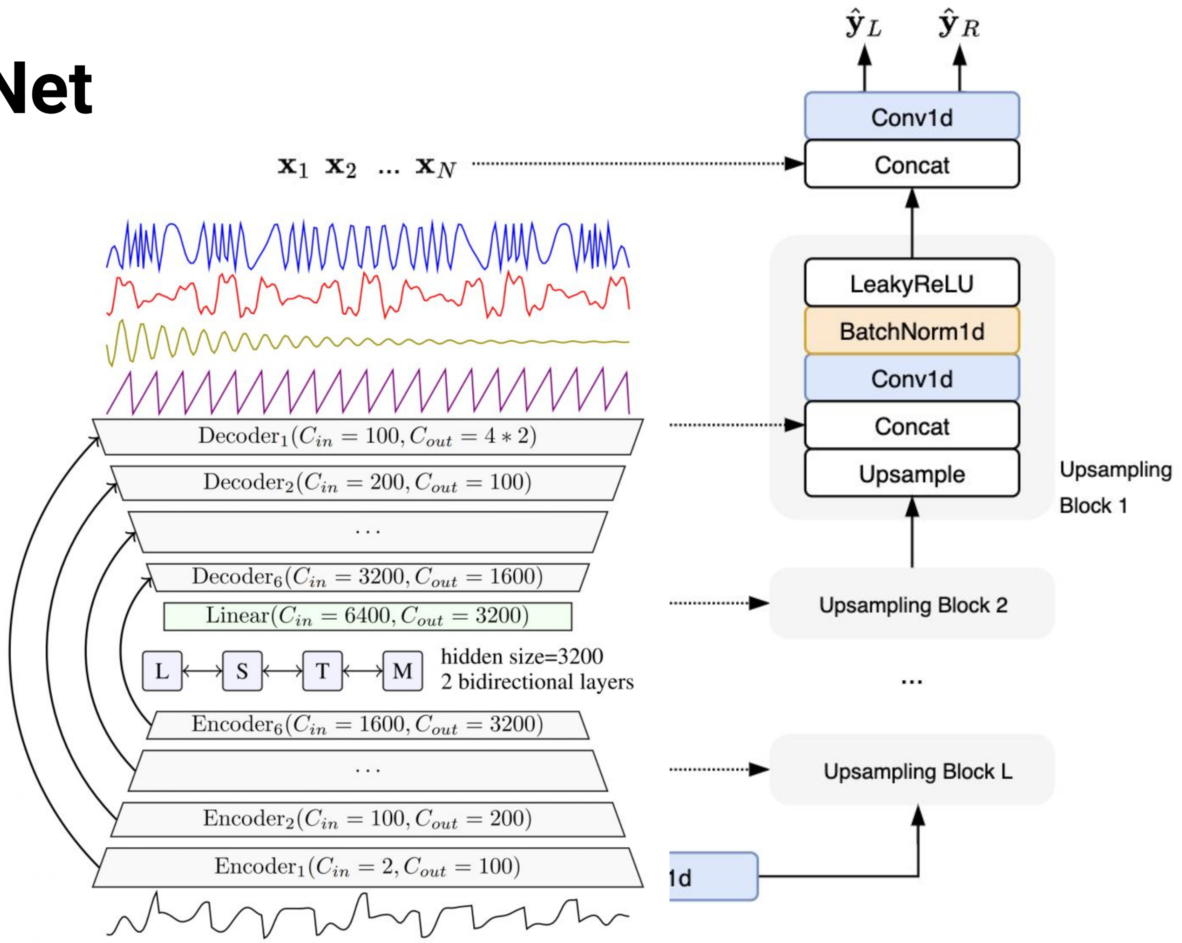


Mix-Wave-U-Net

Upsampling block

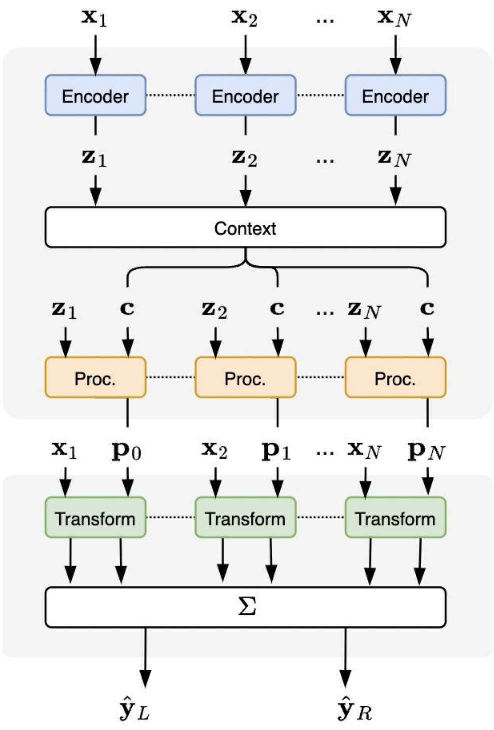


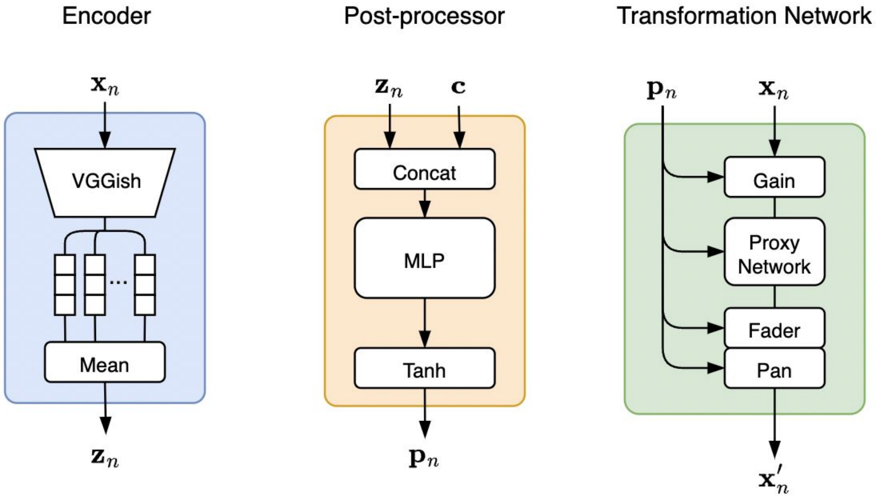
Mix-Wave-U-Net
Output layer



Differentiable Mixing Console

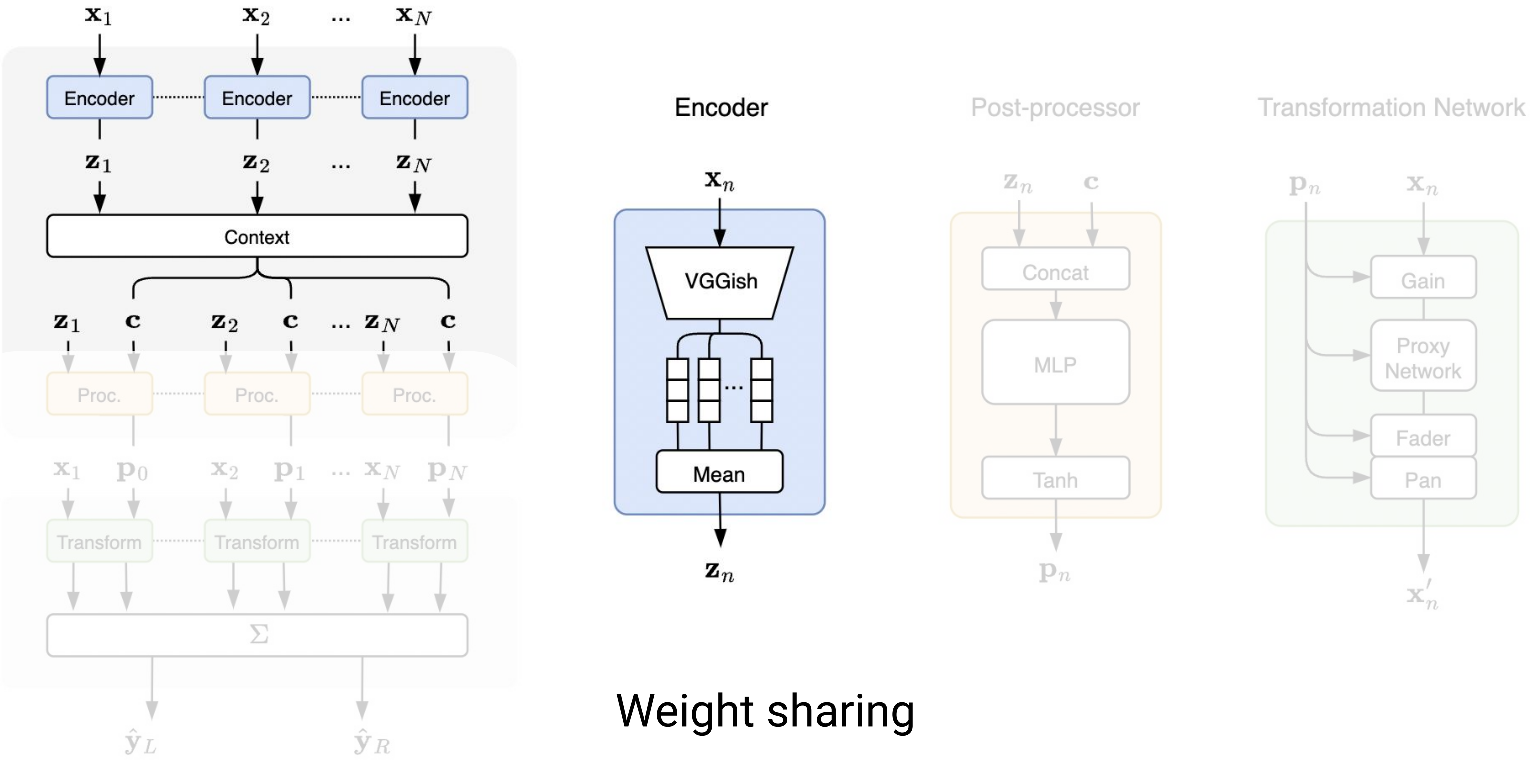
Parameter estimation





Differentiable Mixing Console

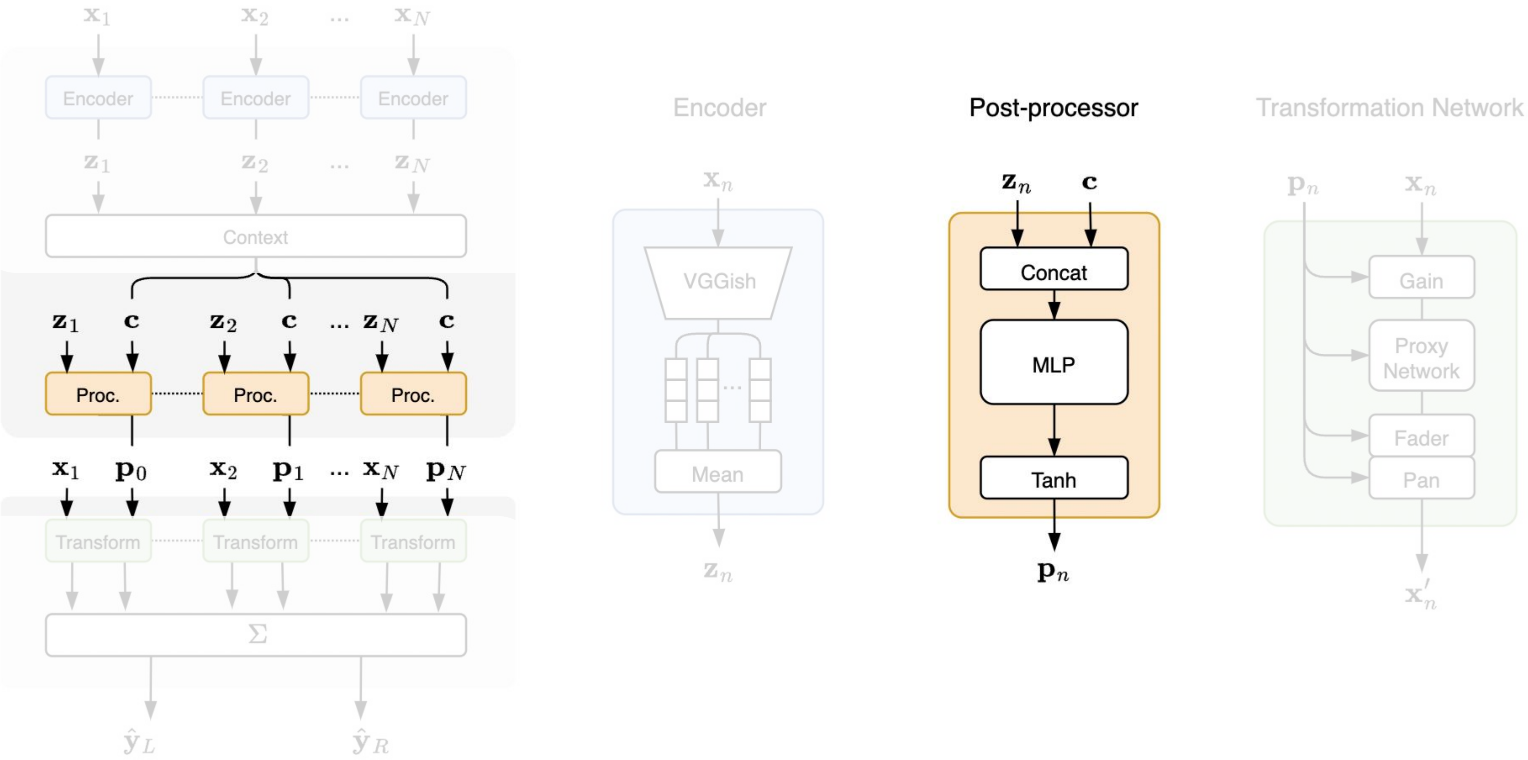
Encoder



Weight sharing

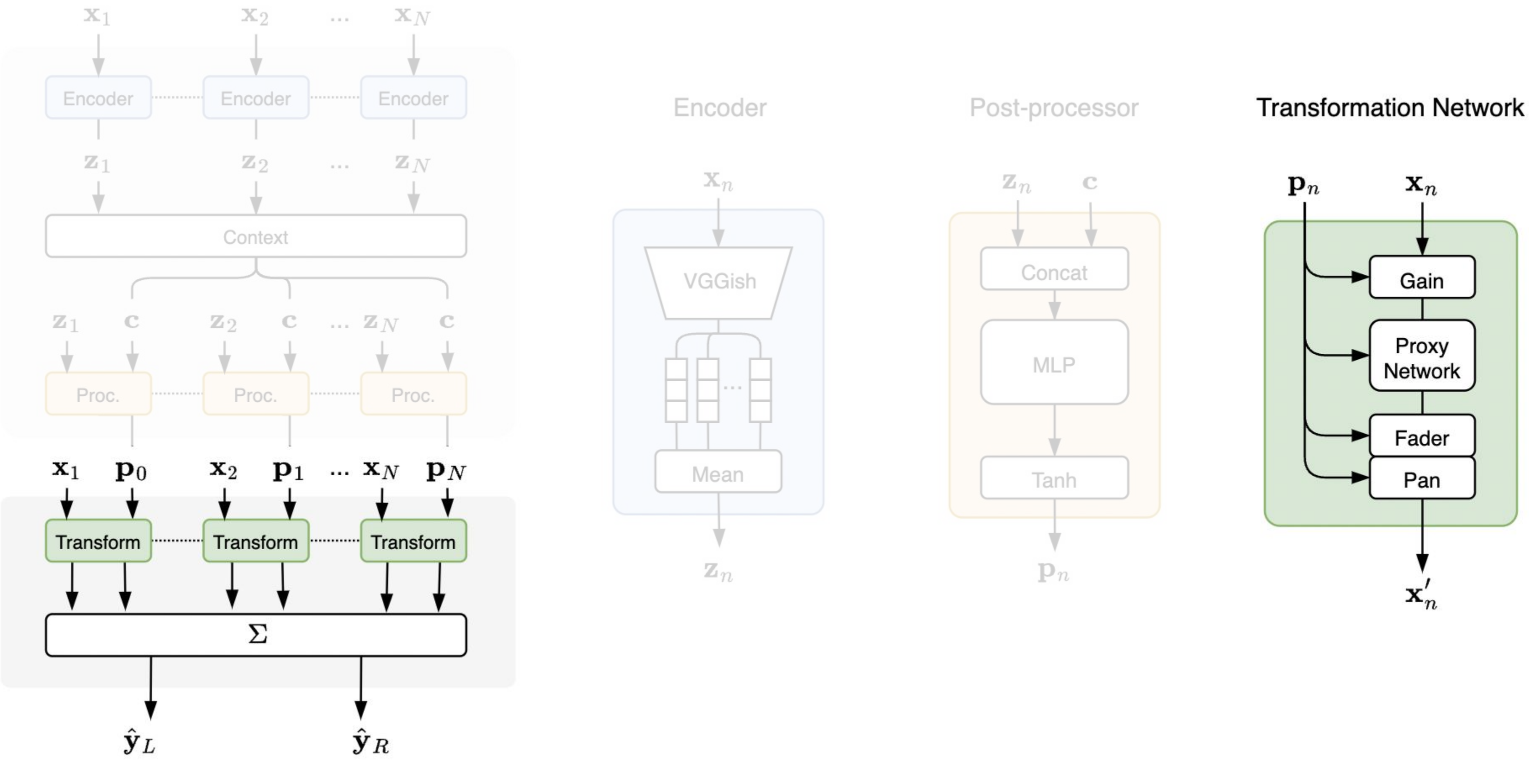
Differentiable Mixing Console

Post-processor



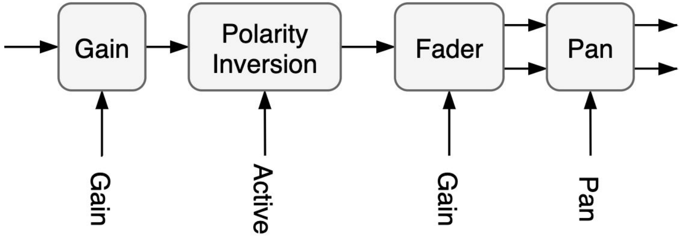
Differentiable Mixing Console

Transformation Network



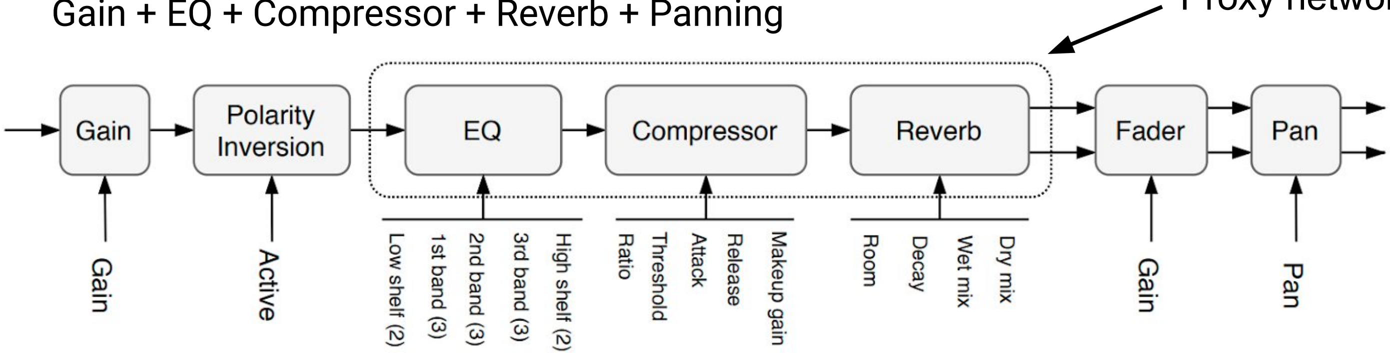
Differentiable Mixing Console

Gain + Panning (Proxy network is not used)



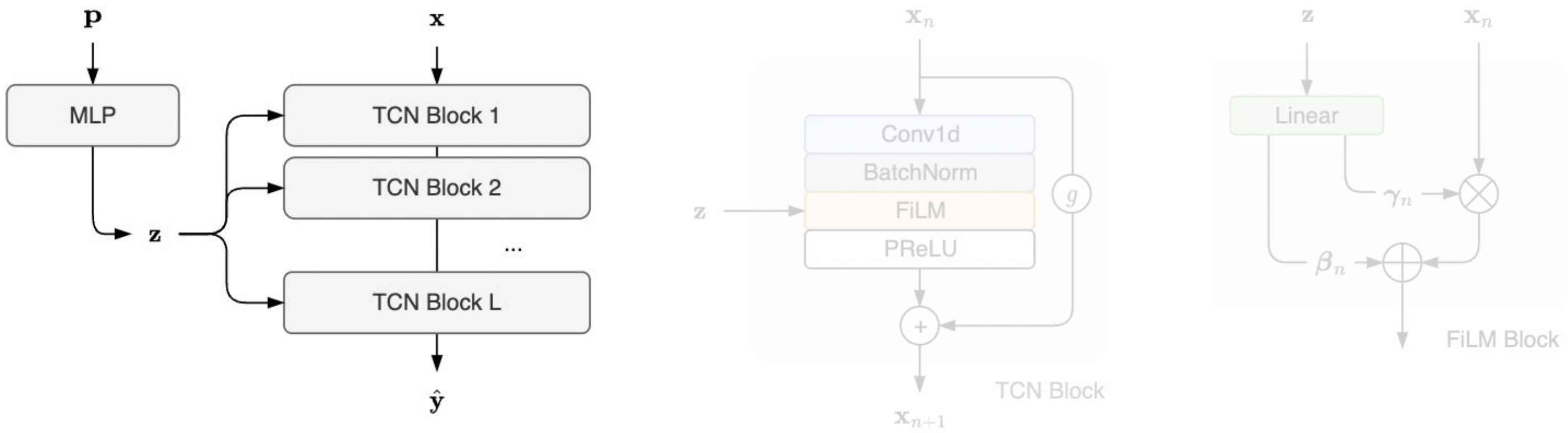
Gain + EQ + Compressor + Reverb + Panning

Proxy network



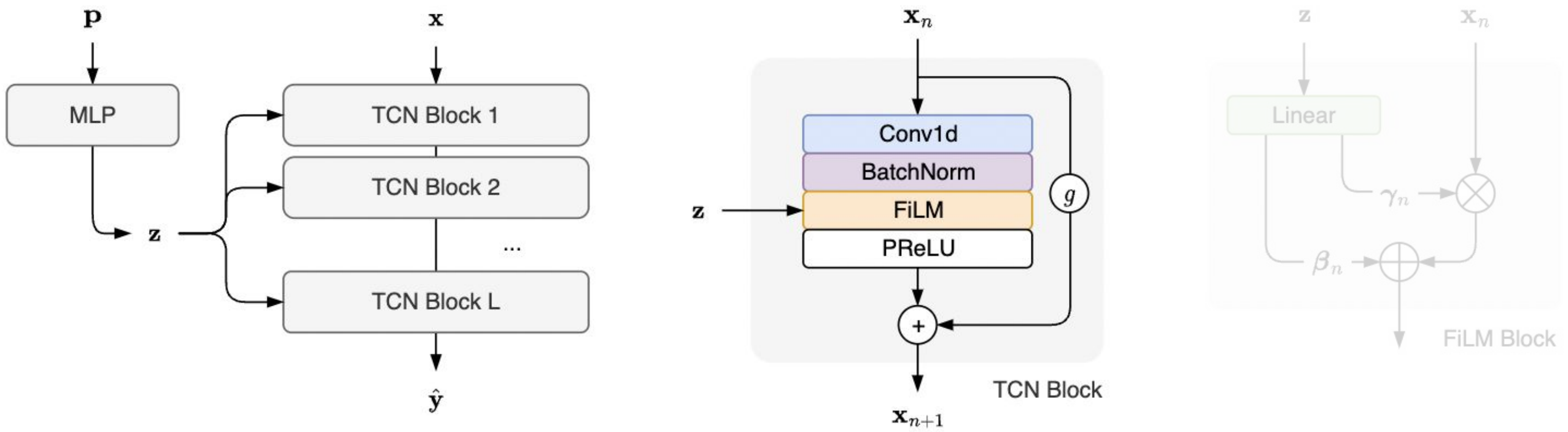
Differentiable Mixing Console

Proxy Networks



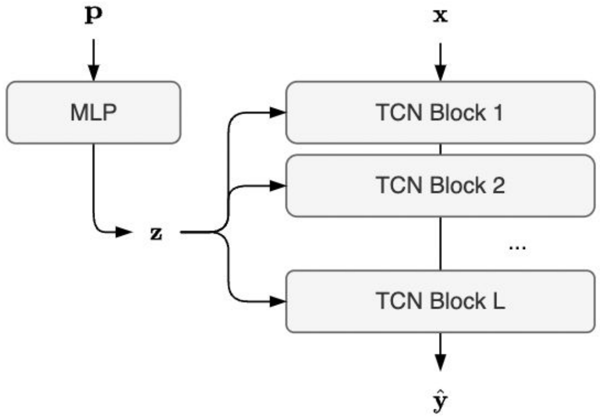
Differentiable Mixing Console

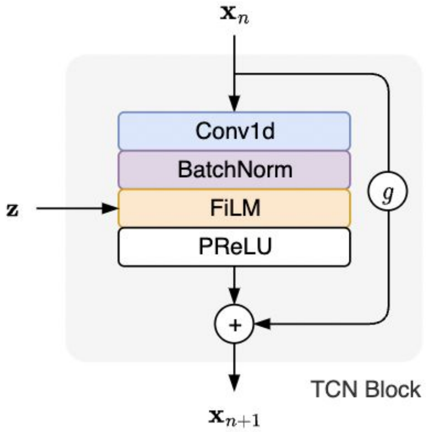
Proxy Networks

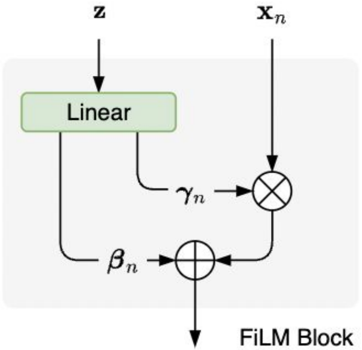


Differentiable Mixing Console

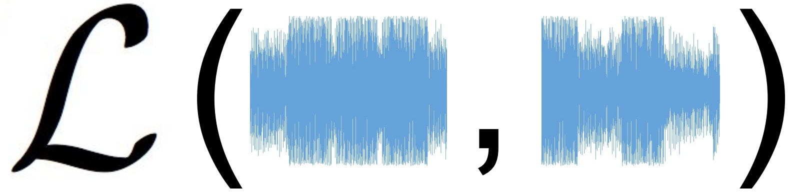
Proxy Networks



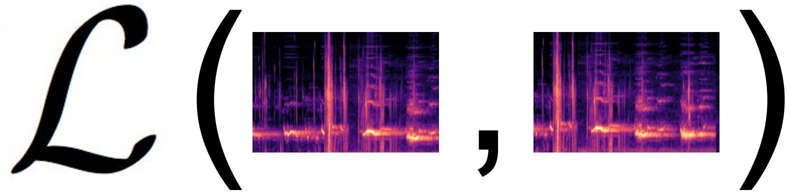




Loss functions



Time domain

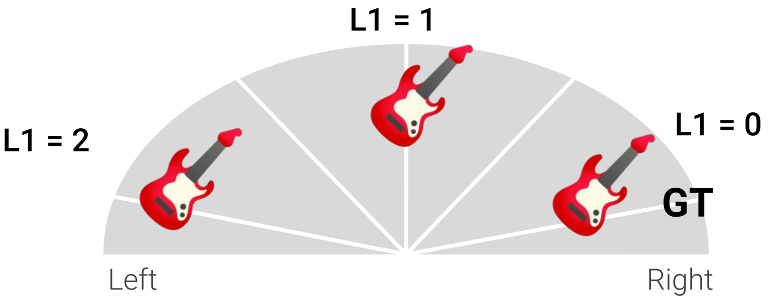


Frequency domain

Stereo loss function

Loss function to encourage realistic mixes

Panning here is more perceptually similar but gives a higher L1 loss



L1 and L2 loss on stereo signals encourage panning all elements to the center.

$$y_{\text{sum}} = y_{\text{left}} + y_{\text{right}}$$

$$y_{\text{diff}} = y_{\text{left}} - y_{\text{right}}$$

$$\ell_{\text{Stereo}}(\hat{y}, y) = \ell_{\text{MR-STFT}}(\hat{y}_{\text{sum}}, y_{\text{sum}}) + \ell_{\text{MR-STFT}}(\hat{y}_{\text{diff}}, y_{\text{diff}})$$

Achieves invariance to stereo (left-right) orientation

auraloss

A collection of audio-focused loss functions in PyTorch

[PDF]

Setup

```
pip install auraloss
```

Usage

```
import torch
import auraloss

mrstft = auraloss.freq.MultiResolutionSTFTLoss()

input = torch.rand(8,1,44100)
target = torch.rand(8,1,44100)

loss = mrstft(input, target)
```

https://github.com/csteinmetz1/auraloss

Loss function	Interface	Reference
Time domain		
Error-to-signal ratio (ESR)	`auraloss.time.ESRLoss()`	Wright & Välimäki, 2019
DC error (DC)	`auraloss.time.DCLoss()`	Wright & Välimäki, 2019
Log hyperbolic cosine (Log-cosh)	`auraloss.time.LogCoshLoss()`	Chen et al., 2019
Signal-to-noise ratio (SNR)	`auraloss.time.SNRLoss()`	
Scale-invariant signal-to-distortion ratio (SI-SDR)	`auraloss.time.SISDRLoss()`	Le Roux et al., 2018
Scale-dependent signal-to-distortion ratio (SD-SDR)	`auraloss.time.SDSDRLoss()`	Le Roux et al., 2018
Frequency domain		
Aggregate STFT	`auraloss.freq.STFTLoss()`	Arik et al., 2018
Aggregate Mel-scaled STFT	`auraloss.freq.MelSTFTLoss(sample_rate)`	
Multi-resolution STFT	`auraloss.freq.MultiResolutionSTFTLoss()`	Yamamoto et al., 2019*
Random-resolution STFT	`auraloss.freq.RandomResolutionSTFTLoss()`	Steinmetz & Reiss, 2020
Sum and difference STFT loss	`auraloss.freq.SumAndDifferenceSTFTLoss()`	Steinmetz et al., 2020
Perceptual transforms		
Sum and difference signal transform	`auraloss.perceptual.SumAndDifference()`	
FIR pre-emphasis filters	`auraloss.perceptual.FIRFilter()`	Wright & Välimäki, 2019

Questions

Break

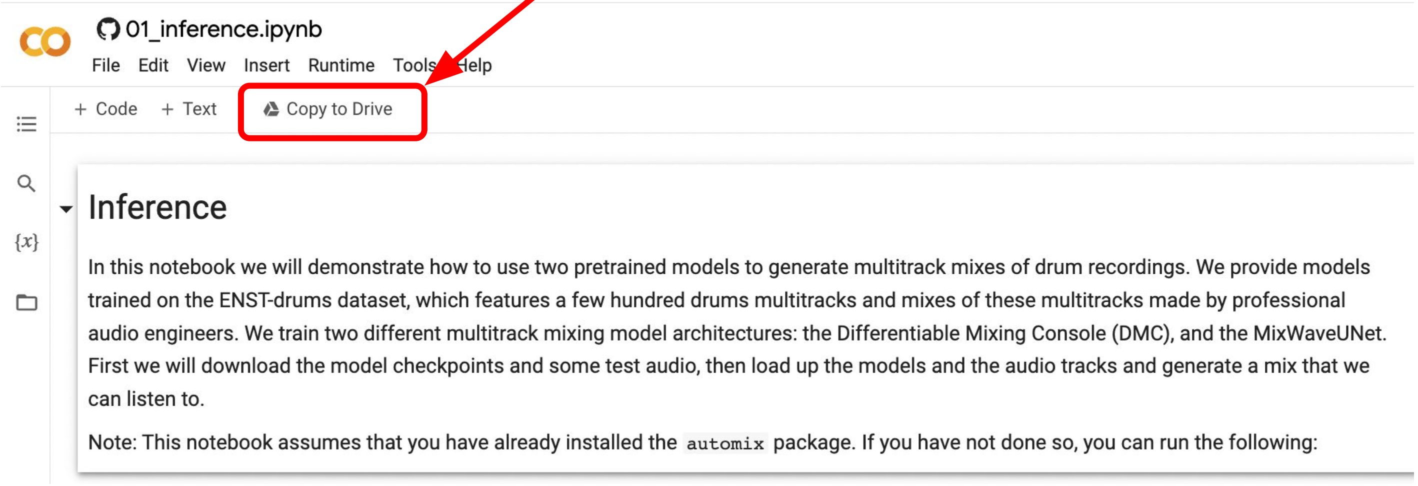
(15min)

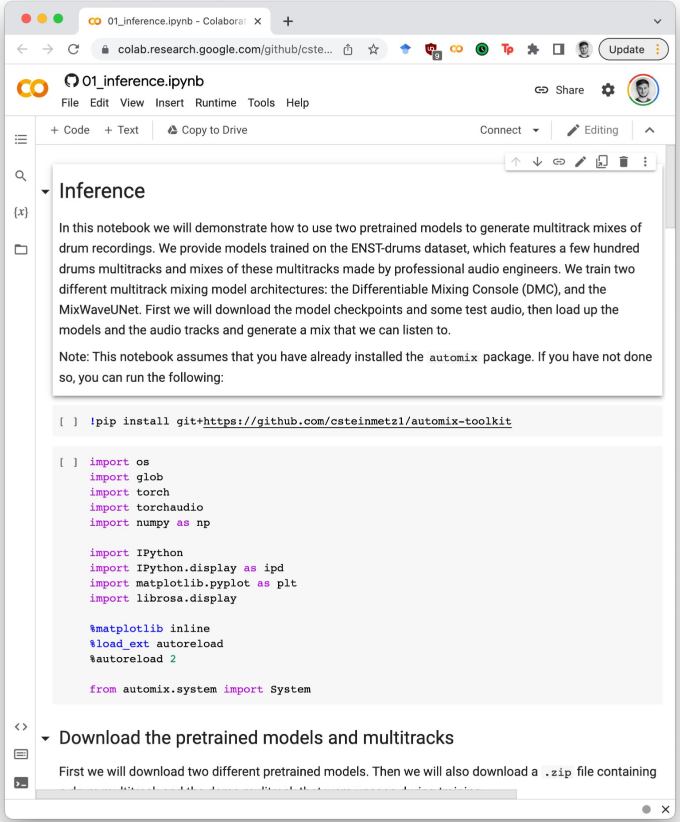
Implementation

Part 3

Soumya Sai Vanka

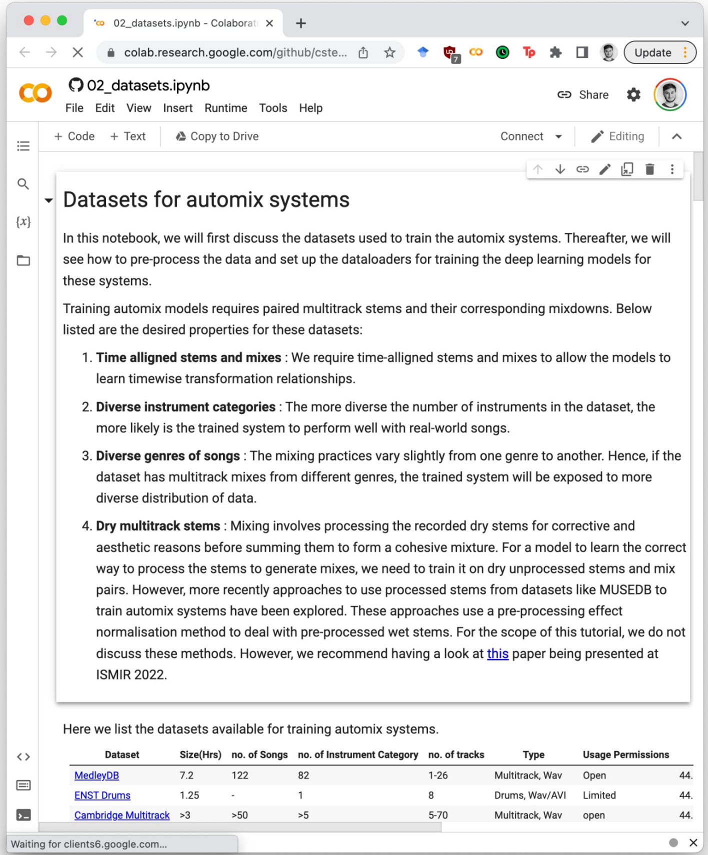
You can save your results and come back later if you click “Copy to Drive”





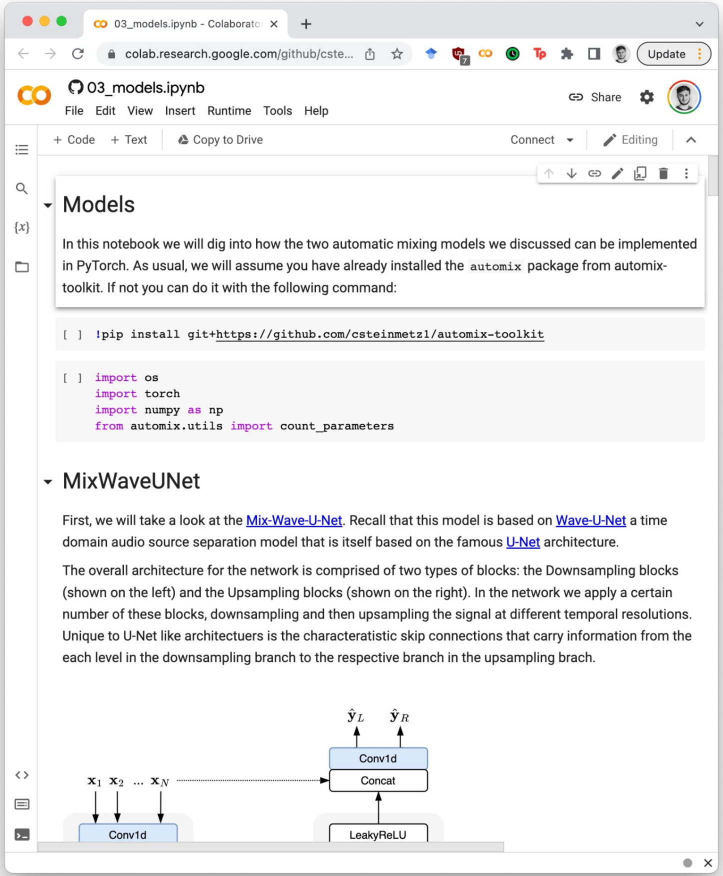
Inference

[Link]



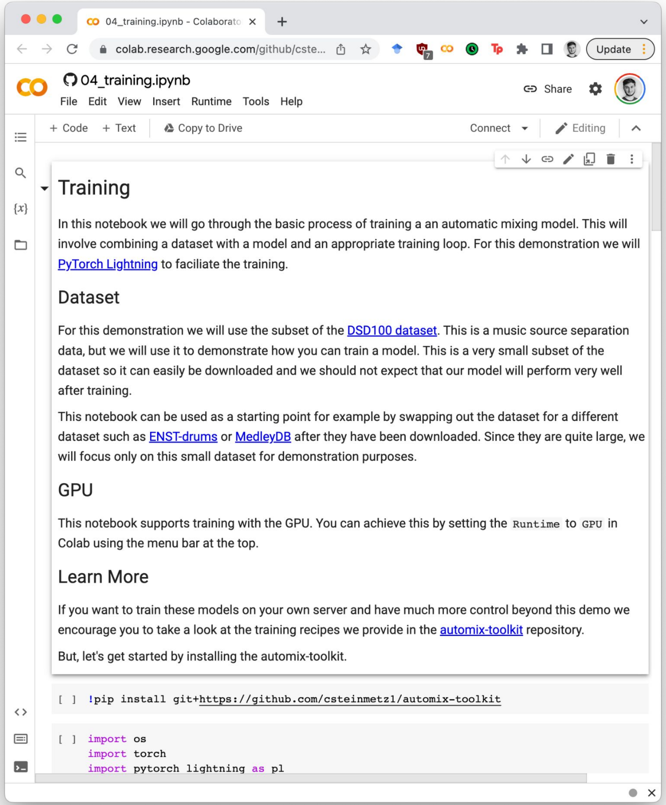
Datasets

[Link](#)


03_models.ipynb - Colaborato
colab.research.google.com/github/cste...
Update
03_models.ipynb
File Edit View Insert Runtime Tools Help
Share
+ Code + Text
Copy to Drive
Connect
Editing
Models
In this notebook we will dig into how the two automatic mixing models we discussed can be implemented in PyTorch. As usual, we will assume you have already installed the automix package from automix-toolkit. If not you can do it with the following command:
[ ] !pip install git+https://github.com/csteinmetz1/automix-toolkit
[ ] import os
import torch
import numpy as np
from automix.utils import count_parameters
MixWaveUNet
First, we will take a look at the Mix-Wave-U-Net. Recall that this model is based on Wave-U-Net a time domain audio source separation model that is itself based on the famous U-Net architecture.
The overall architecture for the network is comprised of two types of blocks: the Downsampling blocks (shown on the left) and the Upsampling blocks (shown on the right). In the network we apply a certain number of these blocks, downsampling and then upsampling the signal at different temporal resolutions. Unique to U-Net like architectuers is the characteristic skip connections that carry information from the each level in the downsampling branch to the respective branch in the upsampling brach.
ŷL ŷR
Conv1d
Concat
x1 x2 ... xN
Conv1d
LeakyReLU

Models

[Link]()



Training

[Link]()

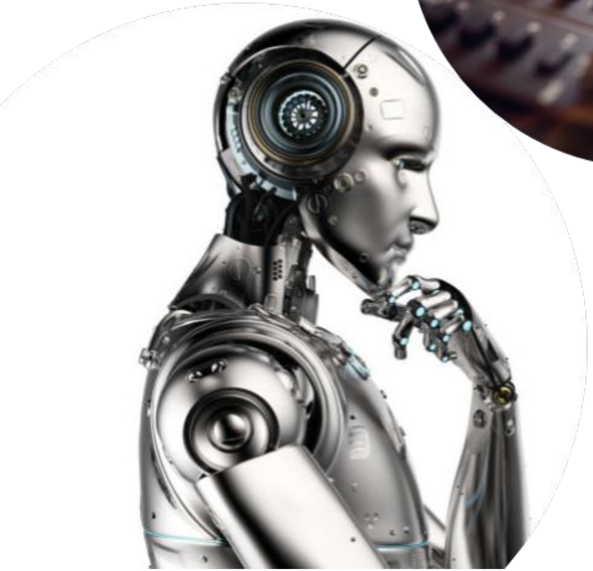
Questions

Evaluation

Part 4

Marco A. Martínez-Ramírez

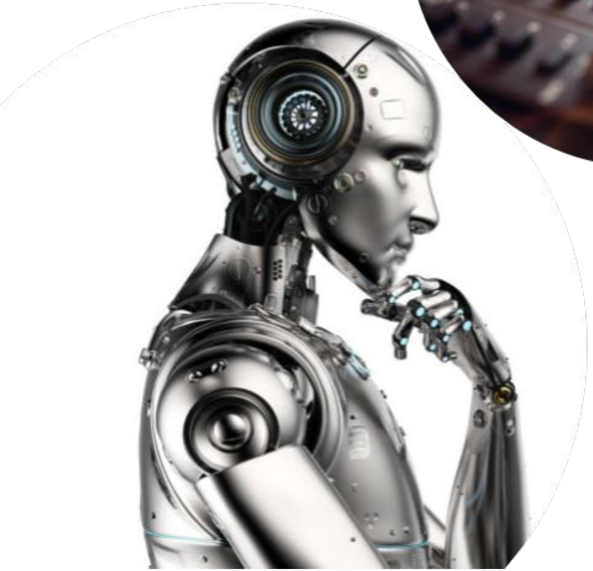
Evaluation



Music mixing is inherently a creative process and therefore a highly subjective task

It cannot be categorized as correct or incorrect

Evaluation



There is not a single metric that will fully encompass the production quality of a generated mix

The use of a professional mix as the ground truth can be an indicator of performance

However, a mix that deviates from the ground truth is not always an aesthetically unpleasant or “bad” mix.

Objective Metrics

- **Objective evaluation of music production tasks remains an open field of research**

- No audio feature, loss function or deep learning embedding have yet
  been found that fully represent solely the mixing processing

- We can use audio features related to mixing audio effects as a way to numerically approximate the evaluation of mixes

Objective Metrics

- **Objective evaluation of music production tasks remains an open field of research**

- No audio feature, loss function or deep learning embedding have yet been found that fully represent solely the mixing processing

- We can use audio features related to mixing audio effects as a way to numerically approximate the evaluation of mixes

Shortcomings

- Cannot capture production quality or aesthetic improvements

- Cannot evidence artifacts within the mix

- Ill-posed problem; deviating from the ground truth does not always mean the mix is incorrect

Audio Features

Spectral features

- EQ and reverberation
- Spectral centroid, bandwidth, contrast, flatness, and roll-off

Spatialisation features

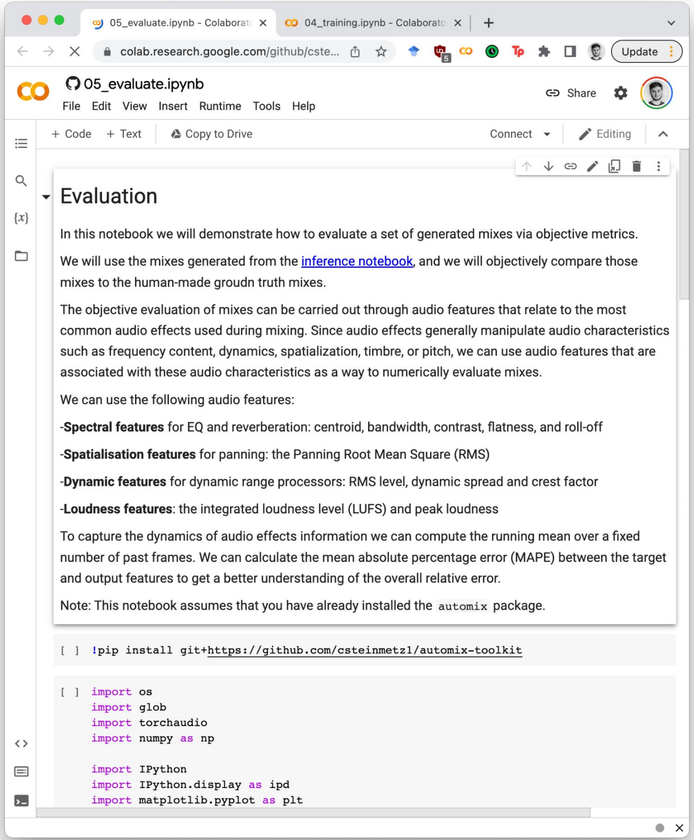
- Panning
- Panning Root Mean Square (RMS)

Dynamic features

- DRC
- RMS level, dynamic spread and crest factor

Loudness features

- The integrated loudness level (LUFS) and peak loudness



Evaluation

[Link]

Listening Test

Perceptual listening tests have become the conventional way to evaluate these systems

There is no standardized test type or platform

We can design tests based on a set of best practices

Adjust them to the specific characteristics of the automatic mixing system

Listening Test

Several design decisions must be taken into account

- Type of test
- Number of stimuli
- Duration of the stimuli
- Criteria to be rated
- Requirements for the participants
- Listening environment

Participants

Preferable to have participants with experience in music mixing, or at least music making or critical listening activities

Participants without such experience are likely to not perceive production differences between mixes

Listening Environment

The preferable listening setup is a listening room with professional monitors and sound installation

If this is not available, the use of high-quality headphones is preferred

Take into account headphones stereo image effect ("inside the head")

Types of test

- **Multi-stimuli tests are often preferred over pairwise or single stimulus tests**
- It is preferable for Participants to focus on the contrasting mix properties between mixes
- Pairwise tests are less reliable and discriminatory when the number of mixes to be compared increases

Types of test

- **Most common types of multi-stimuli test:**

- Multiple Stimuli with Hidden Reference and Anchor (**MUSHRA**) test (ITU-R, 2015)

- Audio Perceptual Evaluation (**APE**) test (De Man and Reiss, 2014)

MUSHRA

- Initially designed for measuring the perceptual quality of audio codecs
- Design constraints represent several limitations when evaluating music mixes

MUSHRA

- Initially designed for measuring the perceptual quality of audio codecs

- Design constraints represent several limitations when evaluating music mixes

Professional human-made mix as reference can be problematic

- Not always rated highly
- Not recommended when the stimuli can outperform the hidden anchor
- Mixes are often not tested for their similarity to a reference mix

MUSHRA

Low and mid anchors

- When participants are experts, it might have a negative impact on the test results
- **Compresses the ratings of the other stimuli**
- Distracts participants from focusing on the contrastive differences within the mixtures
- Not using anchors decreases the number of stimuli, thus, reducing listening time

MUSHRA

Low and mid anchors

- When participants are experts, it might have a negative impact on the test results
- Compresses the ratings of the other stimuli
- Distracts participants from focusing on the contrastive differences within the mixtures
- Not using anchors decreases the number of stimuli, thus, reducing listening time
- **If participants are not experts, the use of low and mid anchors can be beneficial**

MUSHRA

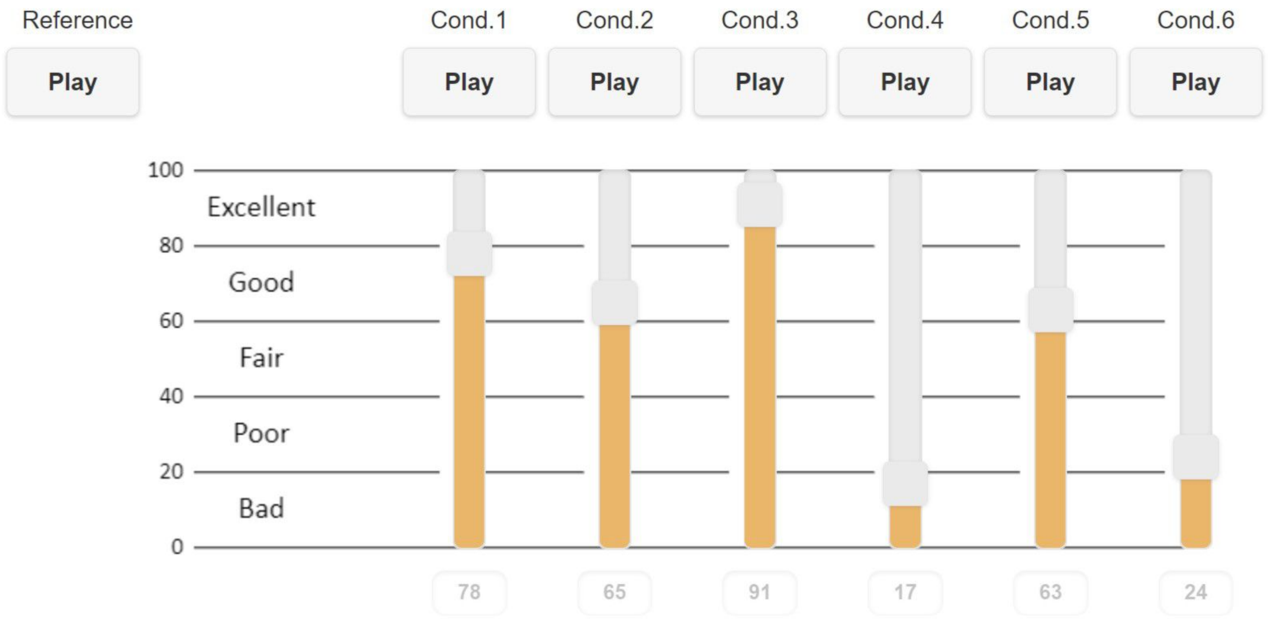
Duration stimuli

- MUSHRA method recommends using stimuli of less than 12 seconds
- Experts consider this duration to be too short to adequately assess quality within a set of mixtures

MUSHRA

In general, it is not recommended to fully follow the MUSHRA methodology, however, this method could be further modified to fit the specific needs of this task

MUSHRA



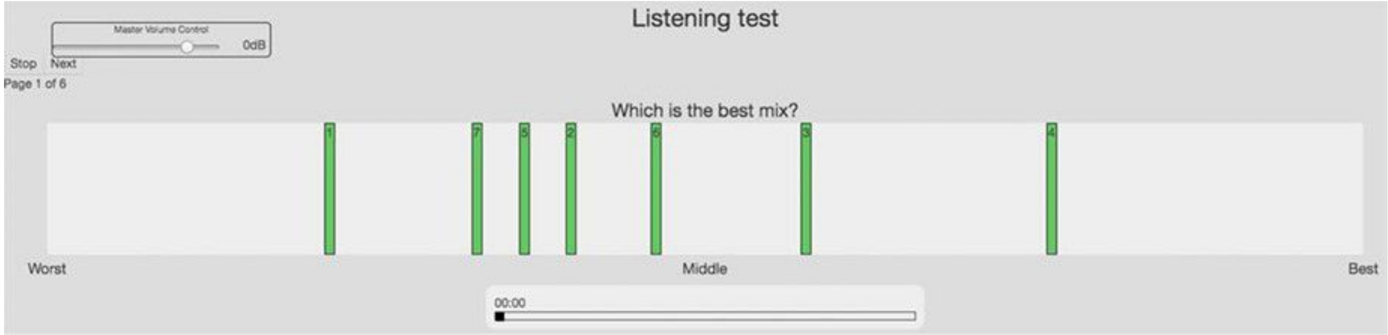
MUSHRA test implemented with webMUSHRA (Schoeffler et al., 2018)

APE

As an alternative for multi-stimuli testing (De Man and Reiss, 2014)

- All the stimuli are placed under the same continuous horizontal line, thus allowing an instant visualization of the ratings

- The use of reference and anchor is optional as well as the maximum length of the stimuli



APE test from Martínez-Ramírez et al. (2021a)

Criteria

- The most common is to ask participants to rate mixes according to their **preference**
- This encompasses both technical and subjective criteria
- Based on a scale from 0 to 1 or from 0 to 100
- With or without the use of semantic labels

Criteria

For a more detailed and discriminatory perceptual ratings, the overall preference could be divided into:

- **Production Value, Clarity and Excitement** (Pestana and Reiss, 2014)
- Preference related to each audio effect, e.g. EQ, reverberation, panning, DRC and overall mixing

Criteria

Production Value

- Technical quality of the mix
- Subjective preferences related to the overall technical quality of the mix
- Considering all the audio mixing characteristics; such as dynamics, EQ, stereo image

Clarity

- Ability to differentiate musical sources
- This is entirely objective
- Corresponds to the perceived masking

Excitement

- A non-technical subjective reaction to the mix
- Not related to an evaluation of quality, but to a more personal perception of novelty
- Considering engaging, intriguing or thought provoking aspects within the mix

Advice

Advice

- **Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.**

- Randomize the order of the stimuli and mixtures to be tested

- Participants with experience in mixing are preferable

- Conduct a pilot listening test

- Always write detailed instructions and, if possible, also provide verbal instructions

- Exclude participants if their total testing time is too short or if their results largely deviate

Advice

- Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.

- **Randomize the order of the stimuli and mixtures to be tested**

- Participants with experience in mixing are preferable

- Conduct a pilot listening test

- Always write detailed instructions and, if possible, also provide verbal instructions

- Exclude participants if their total testing time is too short or if their results largely deviate

Advice

- Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.
- Randomize the order of the stimuli and mixtures to be tested
- **Participants with experience in mixing are preferable**
- Conduct a pilot listening test
- Always write detailed instructions and, if possible, also provide verbal instructions
- Exclude participants if their total testing time is too short or if their results largely deviate

Advice

- Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.
- Randomize the order of the stimuli and mixtures to be tested
- Participants with experience in mixing are preferable
- **Conduct a pilot listening test**
- Always write detailed instructions and, if possible, also provide verbal instructions
- Exclude participants if their total testing time is too short or if their results largely deviate

Advice

- Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.

- Randomize the order of the stimuli and mixtures to be tested

- Participants with experience in mixing are preferable

- Conduct a pilot listening test

- **Always write detailed instructions and, if possible, also provide verbal instructions**

- Exclude participants if their total testing time is too short or if their results largely deviate

Advice

- Participants should be blind to the stimulus as much as possible. The contrary could lead to a negative bias towards fully automated generated mixes.
- Randomize the order of the stimuli and mixtures to be tested
- Participants with experience in mixing are preferable
- Conduct a pilot listening test
- Always write detailed instructions and, if possible, also provide verbal instructions
- **Exclude participants if their total testing time is too short or if their results largely deviate**

Advice

- **Collect additional data, such as age, gender identity, years of mixing experience, and comments**

- Keep the duration of the listening test under 45 minutes
  - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)

- A training stage may be beneficial to participants

- To fully assess a mix, experts prefer segments between 25 and 60 seconds

- Do not use a reference unless is needed for the specific mixing task

- Low and mid anchors are not to be necessary if participants are experts

Advice

- Collect additional data, such as age, gender identity, years of mixing experience, and comments
- **Keep the duration of the listening test under 45 minutes**
  - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)
- A training stage may be beneficial to participants
- To fully assess a mix, experts prefer segments between 25 and 60 seconds
- Do not use a reference unless is needed for the specific mixing task
- Low and mid anchors are not to be necessary if participants are experts

Advice

- Collect additional data, such as age, gender identity, years of mixing experience, and comments

- Keep the duration of the listening test under 45 minutes
    - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)

- **A training stage may be beneficial to participants**

- To fully assess a mix, experts prefer segments between 25 and 60 seconds

- Do not use a reference unless is needed for the specific mixing task

- Low and mid anchors are not to be necessary if participants are experts

Advice

- Collect additional data, such as age, gender identity, years of mixing experience, and comments
- Keep the duration of the listening test under 45 minutes
  - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)
- A training stage may be beneficial to participants
- **To fully assess a mix, experts prefer segments between 25 and 60 seconds**
- Do not use a reference unless is needed for the specific mixing task
- Low and mid anchors are not to be necessary if participants are experts

Advice

- Collect additional data, such as age, gender identity, years of mixing experience, and comments

- Keep the duration of the listening test under 45 minutes
    - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)

- A training stage may be beneficial to participants

- To fully assess a mix, experts prefer segments between 25 and 60 seconds

- **Do not use a reference unless is needed for the specific mixing task**

- Low and mid anchors are not to be necessary if participants are experts

Advice

- Collect additional data, such as age, gender identity, years of mixing experience, and comments

- Keep the duration of the listening test under 45 minutes
    - The max duration without listening fatigue affecting the results is 90 mins (Schatz et al., 2012)

- A training stage may be beneficial to participants

- To fully assess a mix, experts prefer segments between 25 and 60 seconds

- Do not use a reference unless is needed for the specific mixing task

- **Low and mid anchors are not to be necessary if participants are experts**

Advice

- **The number of stimuli per multi-stimulus test page must be less than 12** (Stables et al., 2019)

- If labels are assigned to the rating scale, they must be properly defined and explained to the participants

- Participants prefer synchronized playback between stimuli

- Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)

- Participants must limit the times they adjust the volume of their listening settings

- Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations

Advice

- The number of stimuli per multi-stimulus test page must be less than 12 (Stables et al., 2019)

- **If labels are assigned to the rating scale, they must be properly defined and explained to the participants**

- Participants prefer synchronized playback between stimuli

- Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)

- Participants must limit the times they adjust the volume of their listening settings

- Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations

Advice

- The number of stimuli per multi-stimulus test page must be less than 12 (Stables et al., 2019)

- If labels are assigned to the rating scale, they must be properly defined and explained to the participants

- **Participants prefer synchronized playback between stimuli**

- Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)

- Participants must limit the times they adjust the volume of their listening settings

- Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations

Advice

- The number of stimuli per multi-stimulus test page must be less than 12 (Stables et al., 2019)
- If labels are assigned to the rating scale, they must be properly defined and explained to the participants
- Participants prefer synchronized playback between stimuli
- **Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)**
- Participants must limit the times they adjust the volume of their listening settings
- Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations

Advice

- The number of stimuli per multi-stimulus test page must be less than 12 (Stables et al., 2019)

- If labels are assigned to the rating scale, they must be properly defined and explained to the participants

- Participants prefer synchronized playback between stimuli

- Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)

- **Participants must limit the times they adjust the volume of their listening settings**

- Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations

Advice

- The number of stimuli per multi-stimulus test page must be less than 12 (Stables et al., 2019)

- If labels are assigned to the rating scale, they must be properly defined and explained to the participants

- Participants prefer synchronized playback between stimuli

- Loudness normalize, since loudness should not influence the rated criteria (except for the cases where loudness is crucial to the criteria)

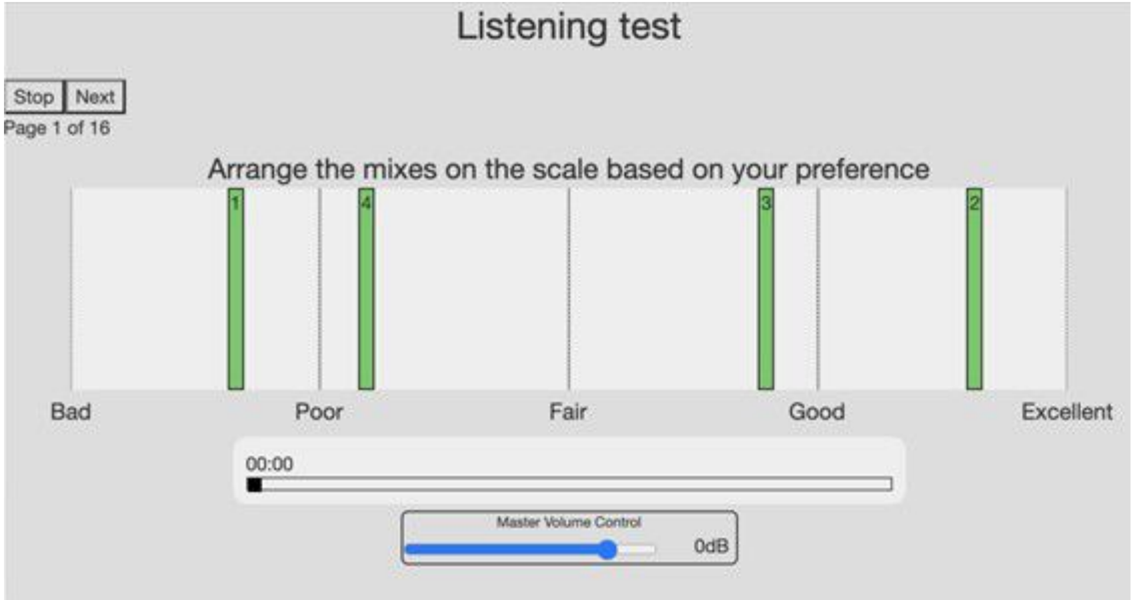
- Participants must limit the times they adjust the volume of their listening settings

- **Professional speakers are preferred. Test exclusively with speakers or headphones, but not allow both listening configurations**

Platforms for multi-stimuli tests

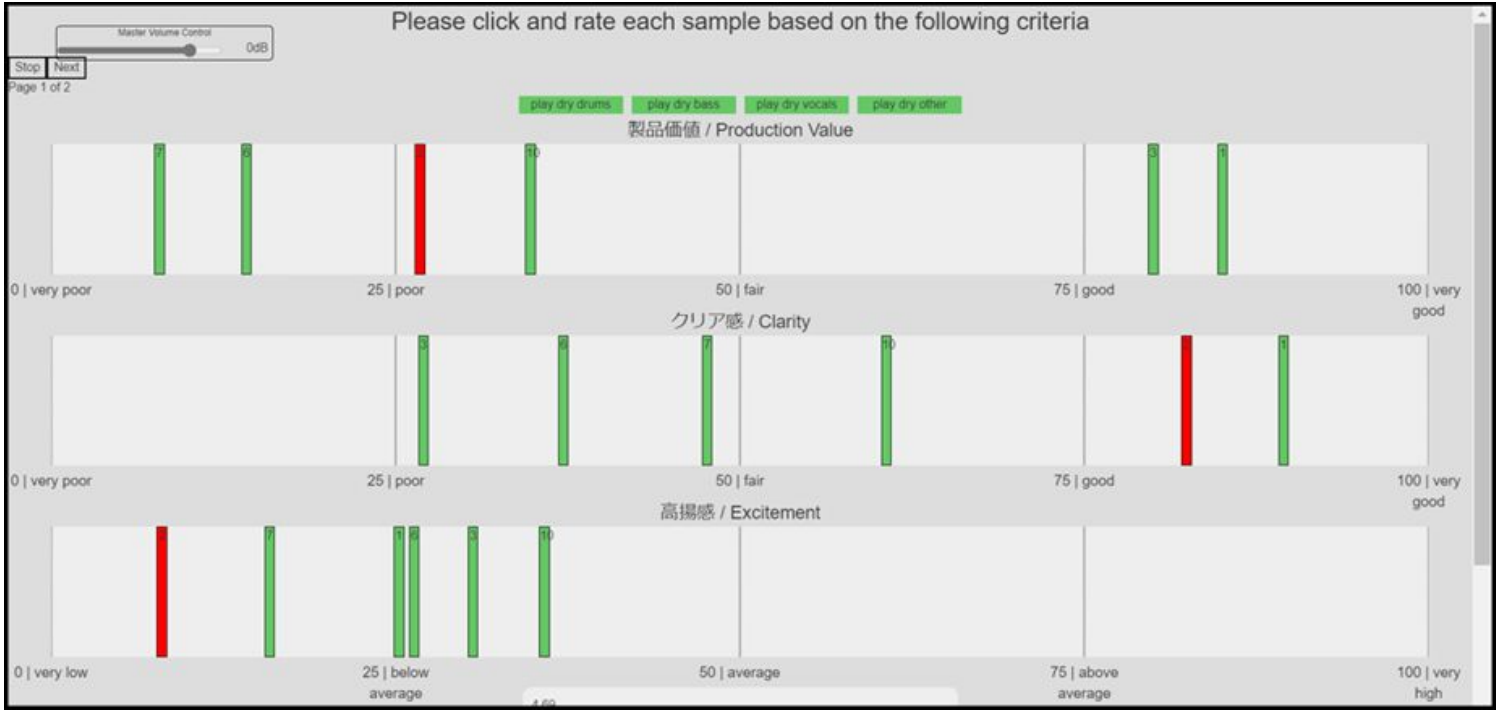
Platform	Multi-stimuli test	Features	Usage
Web Audio Evaluation Tool (Jillings et al., 2015)	-MUSHRA -APE -Discrete -Reference is optional	-Training stage -Loudness normalization -Synchronized playback -Randomization	-Requires server -PHP support has not been updated -Customization with effort
webMUSHRA (Schoeffler et al., 2018)	-MUSHRA	-Training stage -Fade-in/out -Synchronized playback -Randomization	-Requires server -Customization with effort
goListen (Barry et al., 2021b)	-MUSHRA -Reference is optional	-Synchronized playback -Randomization	-Requires account -Does not require server -Customization with effort -Ease-of-use

Platforms



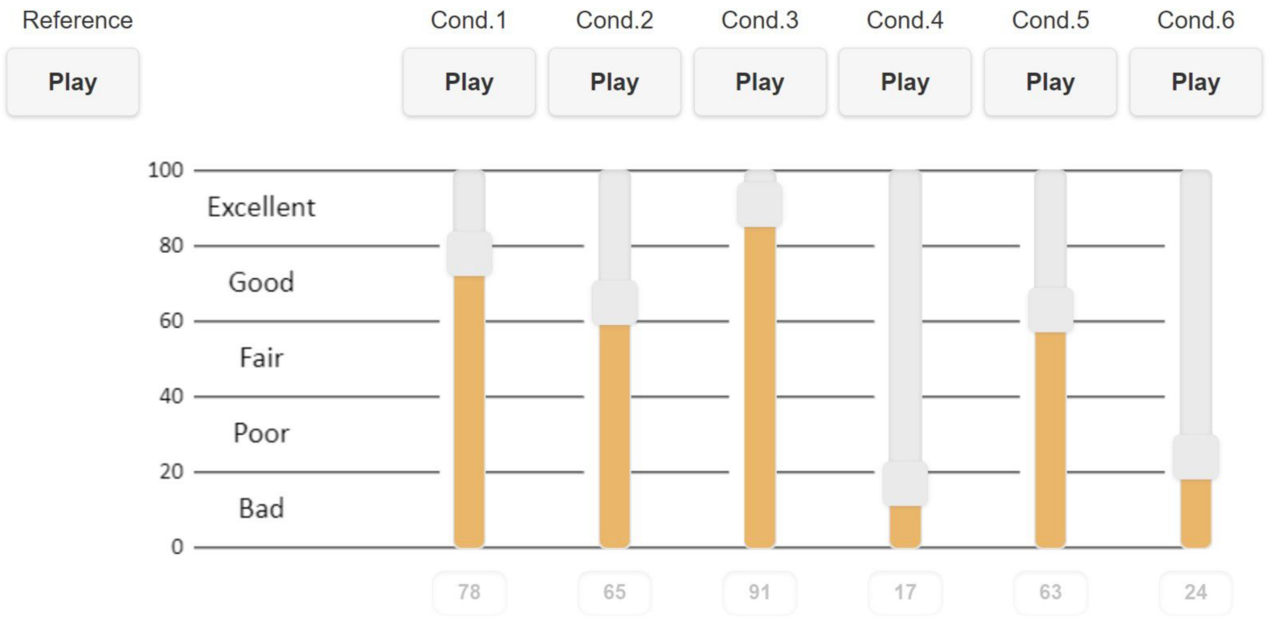
APE test implemented with the Web Audio Evaluation Tool. Test from Steinmetz et al., 2021c

Platforms



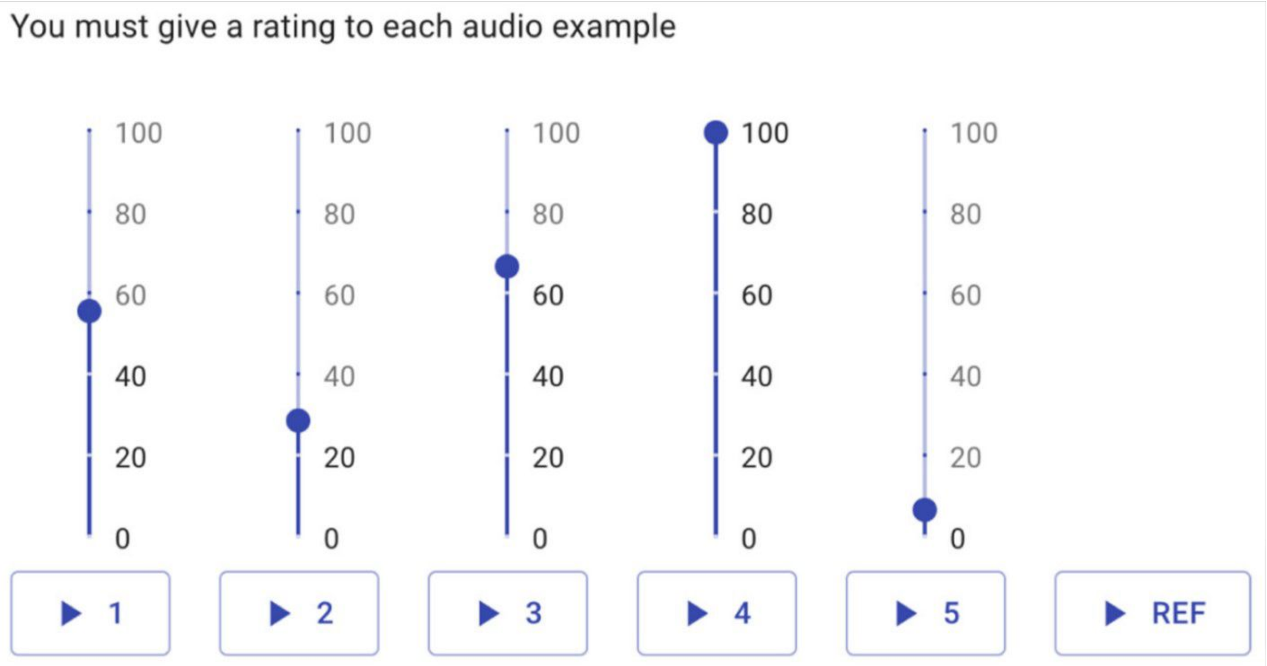
APE test implemented with the Web Audio Evaluation Tool. Test from Martínez-Ramírez et al. (2022). For this test, dry stems were used as references.
This is based on feedback from pilot tests and was proposed by the expert participants

Platforms



MUSHRA test implemented with webMUSHRA (Schoeffler et al., 2018)

Platforms



MUSHRA test implemented with goListen (Barry et al., 2021a)

Listening Test Example

- Please open a listening test example at

  https://golisten.ucd.ie/task/mushra-test/638b0c03d6a905906a2c4402

Listening Test Example

- Please open a listening test example at

  https://golisten.ucd.ie/task/mushra-test/638b0c03d6a905906a2c4402

- Which mix is the best based on your preference ?
- Which one do you think is a human mix (if there is any) ?
- Can you find the low anchor ?

Listening Test Example

- Mix # 1 - (Koo et al., 2022a) - Music Mixing Style Transfer with reference from MUSDB18
- Mix # 2 - Mono mix
- Mix # 3 - Gary’s mix
- Mix # 4 - DMC mix trained with MedleyDB - Gain and Panning
- Mix # 5 - (Martinez-Ramirez et al., 2022) - Trained with MUSDB18
- Mix # 6 - (Martinez-Ramirez et al., 2022) - Trained with large dataset

Listening Test Example

- Mix # 1 - (Koo et al., 2022a) - Music Mixing Style Transfer with reference from MUSDB18
- Mix # 2 - Mono mix
- Mix # 3 - Gary's mix
- Mix # 4 - DMC mix trained with MedleyDB - Gain and Panning
- Mix # 5 - (Martinez-Ramirez et al., 2022) - Trained with MUSDB18
- Mix # 6 - (Martinez-Ramirez et al., 2022) - Trained with large dataset

Song: Isolate - Flare

Future directions

Objective Metrics

- Deep features such as the embedding output of the Fx encoder proposed in (Koo et al., 2022a) could also be used as an indicator of similarity for mixing processing

- Leveraging on general purpose deep features related to audio perception, such as the Fréchet Audio Distance (Kilgour et al., 2019) can also be investigated

Explore limitations of the objective and subjective evaluation methods

- How can we measure whether the generated mixes have long-temporal coherence ?
- To measure mixing style coherence within different song elements such as verses, choruses

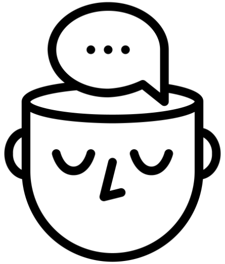
Questions

Conclusion

Part 5

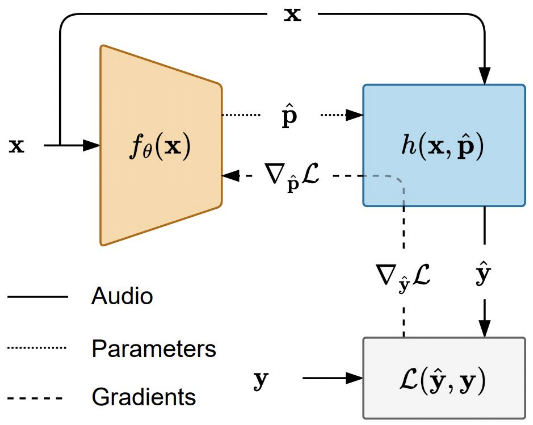
Christian J. Steinmetz

Soumya Sai Vanka



Future Directions

Differentiable Signal Processing



- Controlling audio effects using NN:
  - Neural proxies
  - Gradient approximation methods

- Implementing audio effects as differentiable effects (can be embedded into the neural network pipeline for training and backpropagation)
  - Neural network can learn to control audio effects
  - Implementations available for dynamic range compressor, EQ, Artificial reverberation, and distortion.
  - Differentiable mixing console with the chain of differentiable effects

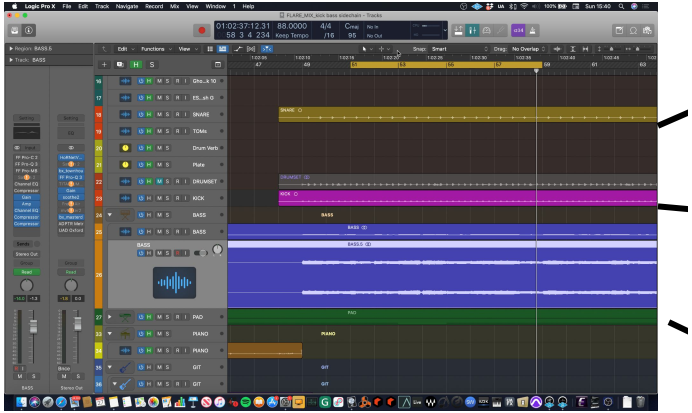
Datasets

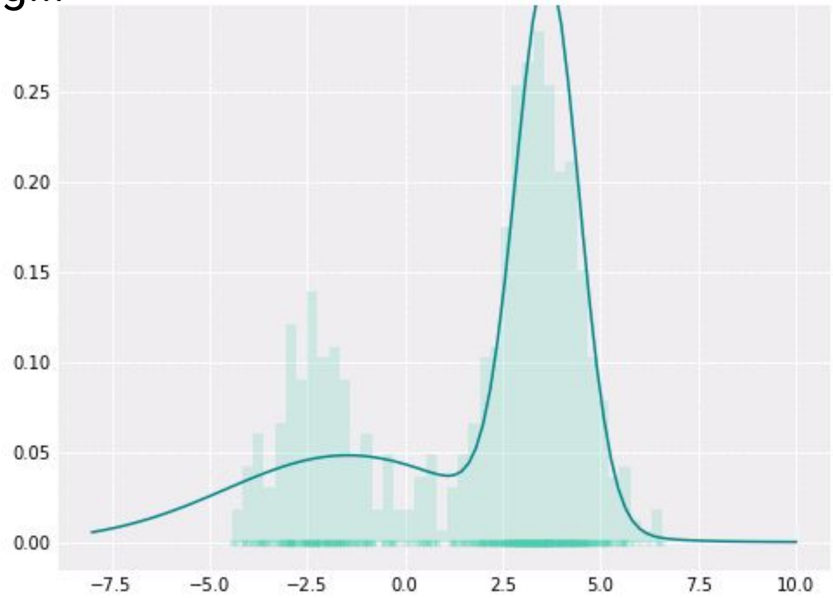
- Ideal: Creating large, annotated, high-quality, open-source multitrack datasets

- Making the best use of what we already have: Can we use Source Separation datasets?
    - Recent work: (by Martinez et. al) uses pre-processing block for audio effect normalisation
    - Utilises source separation datasets for training automix models
    - Next steps: Train Source Separation models to not just separate tracks but also **remove audio effects**; generated dry stems could be used for remixing

Generative models

The mixing task is a one to many mapping...





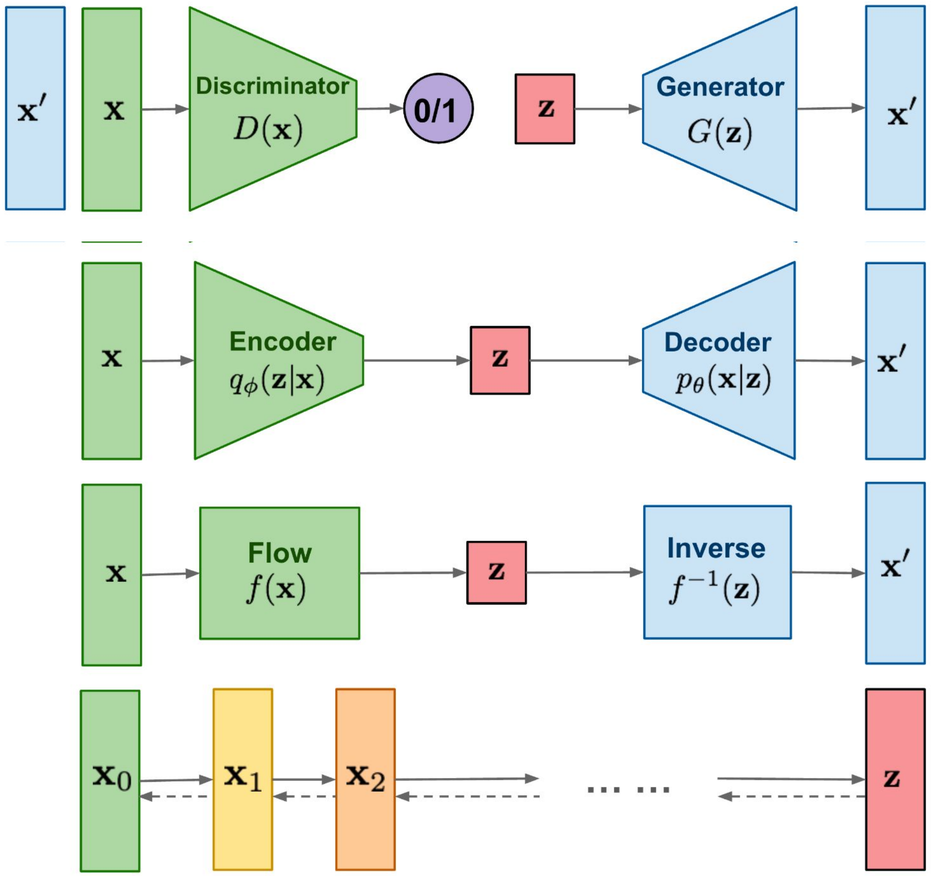
So we should treat it as such.

GAN: Adversarial training

VAE: maximize variational lower bound

Flow-based models: Invertible transform of distributions

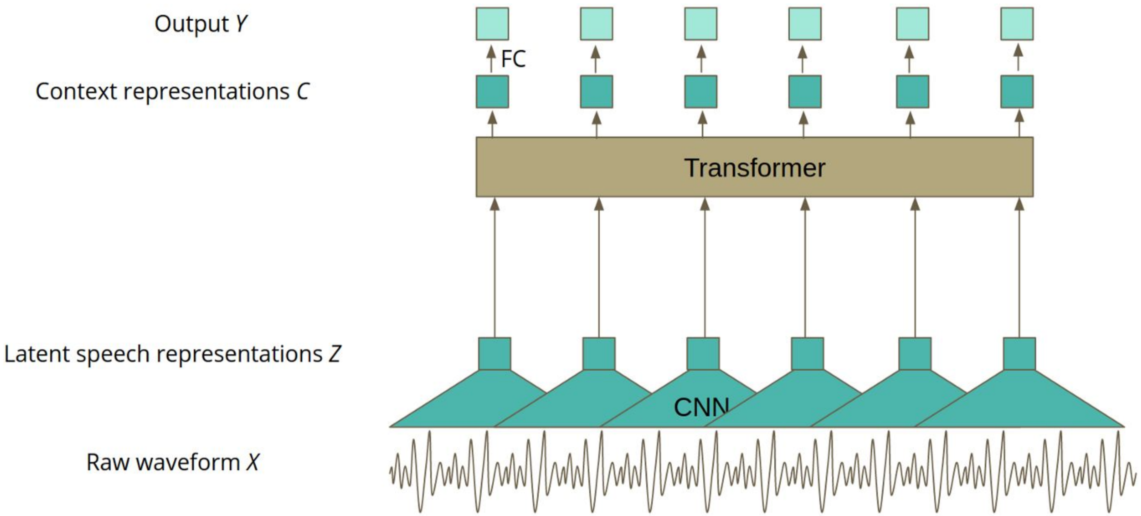
Diffusion models: Gradually add Gaussian noise and then reverse

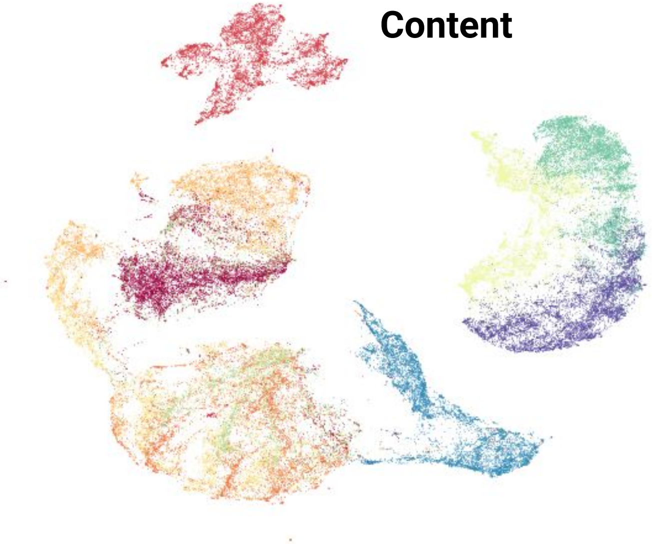


Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log.

Audio Production Representations

wav2vec





Content

Can we build audio reprs. that encode only audio production details?

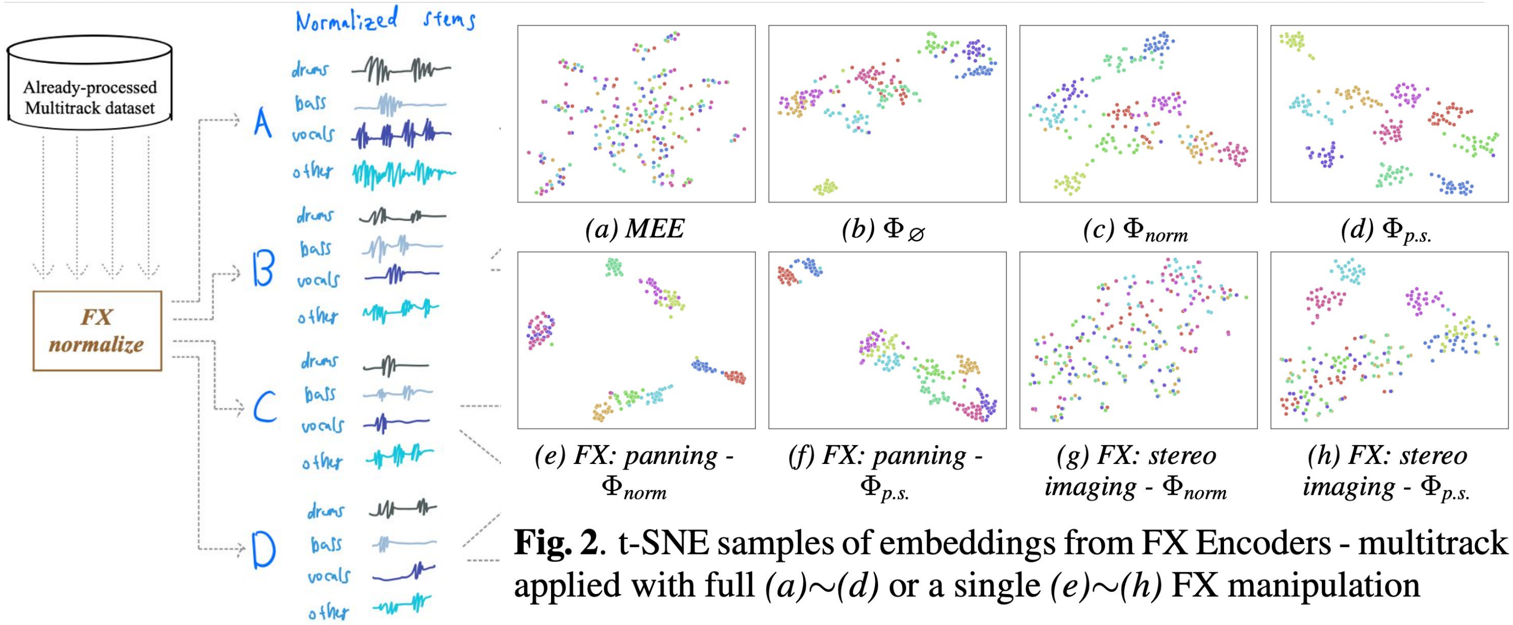


Fig. 2. t-SNE samples of embeddings from FX Encoders - multitrack applied with full *(a)*∼*(d)* or a single *(e)*∼*(h)* FX manipulation

Music Mixing Style Transfer: A Contrastive Learning Approach To Disentangle Audio Effects
Koo et al., arXiv, 2022 https://arxiv.org/abs/2211.02247

Takeaways

1. Mixing is a task that maps creative ideas and emotion to technical parameters
2. Approaches are often either *direct transformation* or *parameter estimation*
3. Evaluation remains challenging and we rely on well design listening tests
4. Many open questions and challenges with potentially fruitful outcomes

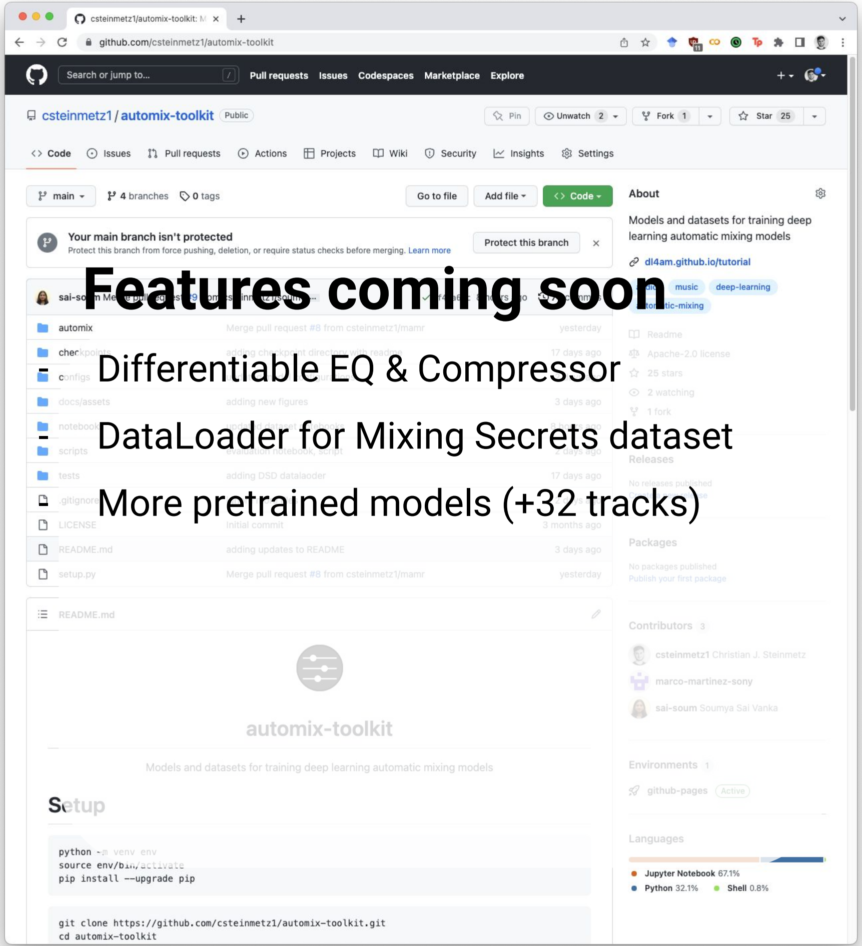
Resources

automix-toolkit

https://github.com/csteinmetz1/automix-toolkit

Star

Star it on GitHub

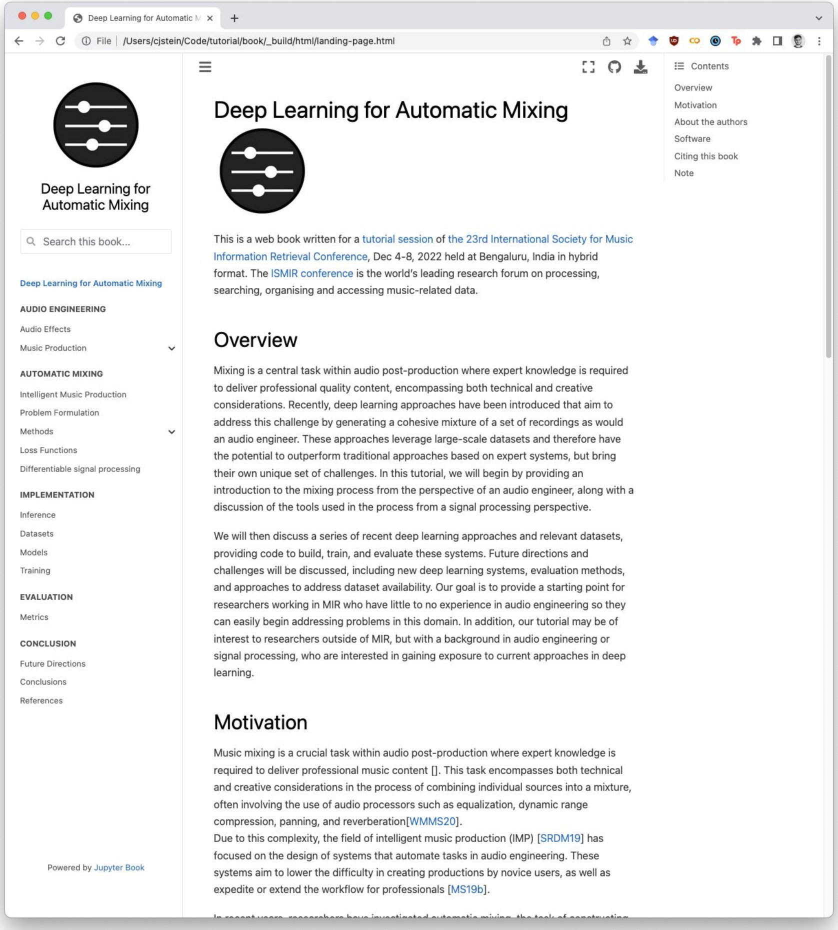


Features coming soon

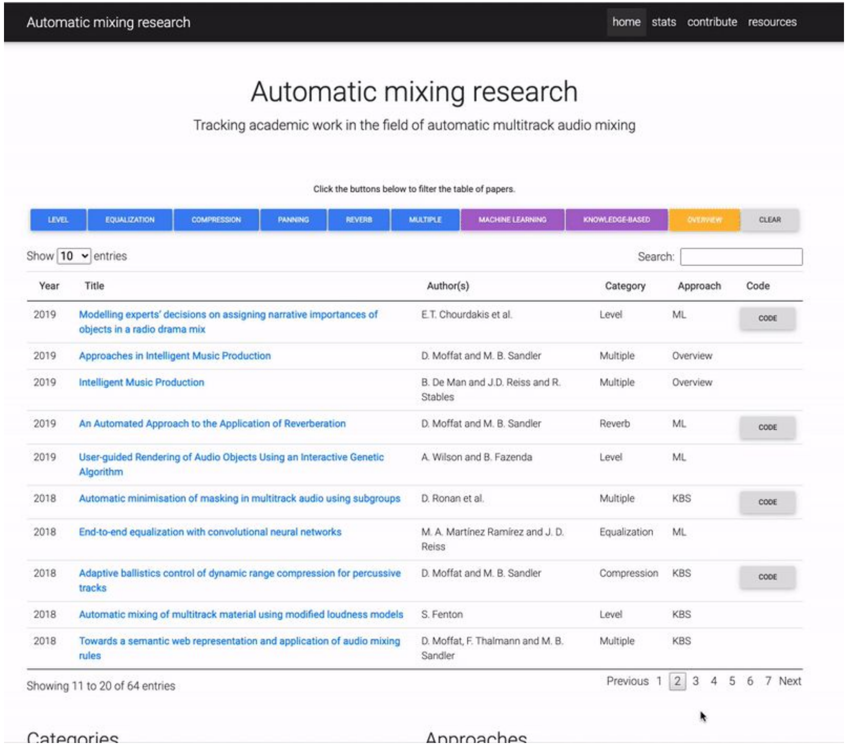
- Differentiable EQ & Compressor
- DataLoader for Mixing Secrets dataset
- More pretrained models (+32 tracks)

Book

https://dl4am.github.io/tutorial



More works on automatic mixing research


Automatic mixing research

home stats contribute resources

Automatic mixing research

Tracking academic work in the field of automatic multitrack audio mixing

Click the buttons below to filter the table of papers.

LEVEL EQUALIZATION COMPRESSION PANNING REVERB MULTIPLE MACHINE LEARNING KNOWLEDGE-BASED OVERVIEW CLEAR

Show 10 entries Search:

Year	Title	Author(s)	Category	Approach	Code
2019	Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix	E.T. Chourdakis et al.	Level	ML	CODE
2019	Approaches in Intelligent Music Production	D. Moffat and M. B. Sandler	Multiple	Overview	
2019	Intelligent Music Production	B. De Man and J.D. Reiss and R. Stables	Multiple	Overview	
2019	An Automated Approach to the Application of Reverberation	D. Moffat and M. B. Sandler	Reverb	ML	CODE
2019	User-guided Rendering of Audio Objects Using an Interactive Genetic Algorithm	A. Wilson and B. Fazenda	Level	ML	
2018	Automatic minimisation of masking in multitrack audio using subgroups	D. Ronan et al.	Multiple	KBS	CODE
2018	End-to-end equalization with convolutional neural networks	M. A. Martínez Ramírez and J. D. Reiss	Equalization	ML	
2018	Adaptive ballistics control of dynamic range compression for percussive tracks	D. Moffat and M. B. Sandler	Compression	KBS	CODE
2018	Automatic mixing of multitrack material using modified loudness models	S. Fenton	Level	KBS	
2018	Towards a semantic web representation and application of audio mixing rules	D. Moffat, F. Thalmann and M. B. Sandler	Multiple	KBS	

Showing 11 to 20 of 64 entries

Previous 1 2 3 4 5 6 7 Next

Categories Approaches

Searchable/filterable table of relevant papers and stats

https://csteinmetz1.github.io/AutomaticMixingPapers

AI x Audio Engineering

Discord Community

https://discord.gg/tPNuUQzR

Citation

```
@book{steinmetz2022automix,
    author = {Steinmetz, Christi
              and Martínez, Marc
    month = {December},
    publisher = {ISMIR},
    title = {Deep Learning for A
    year = 2022,
    url = {https://dl4am.github.
}
```

ISMIR 2022 Tutorial:
Deep learning for automatic mixing

Christian J. Steinmetz[1]
c.j.steinmetz@qmul.ac.uk

Soumya Sai Vanka[1]
s.s.vanka@qmul.ac.uk

Gary Bromham[1]
g.bromham@qmul.ac.uk

Marco A. Martínez Ramírez[2]
marco.martinez@sony.com

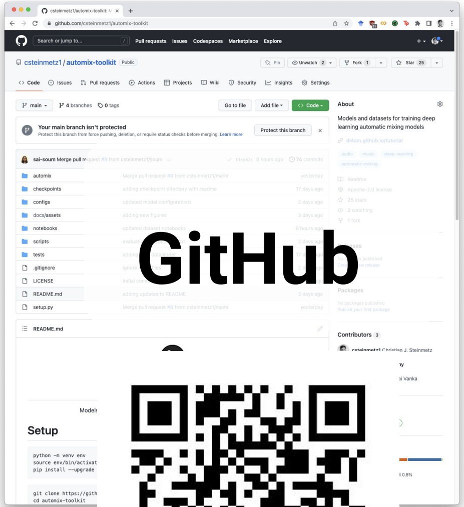
Centre for Digital Music, Queen Mary University of London[1]
Sony Group Corporation[2]

Abstract

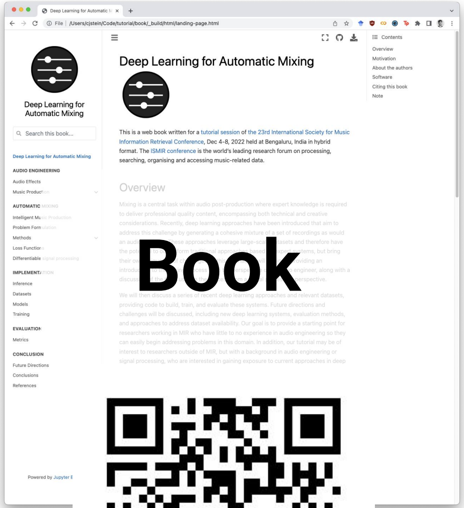
Mixing is a central task within audio post-production where expert knowledge is required to deliver professional quality content, encompassing both technical and creative considerations. Recently, deep learning approaches have been introduced that aim to address this challenge by generating a cohesive mixture of a set of recordings as would an audio engineer. These approaches leverage large-scale datasets and therefore have the potential to outperform traditional approaches based on expert systems, but bring their own unique set of challenges. In this tutorial, we begin by providing an introduction to the mixing process from the perspective of an audio engineer, along with a discussion of the tools used in the process from a signal processing perspective. We then discuss a series of recent deep learning approaches and relevant datasets, providing code to build, train, and evaluate these systems. Future directions and challenges will be discussed, including new deep learning systems, evaluation methods, and approaches to address dataset availability. Our goal is to provide a starting point for researchers working in MIR who have little to no experience in audio engineering so they can easily begin addressing problems in this domain. In addition, our tutorial may be of interest to researchers outside of MIR, but with a background in audio engineering or signal processing, who are interested in gaining exposure to current approaches in deep learning.

On arXiv soon....

Final Questions



GitHub



Book

Discord

Book

https://dl4am.github.io/tutorial